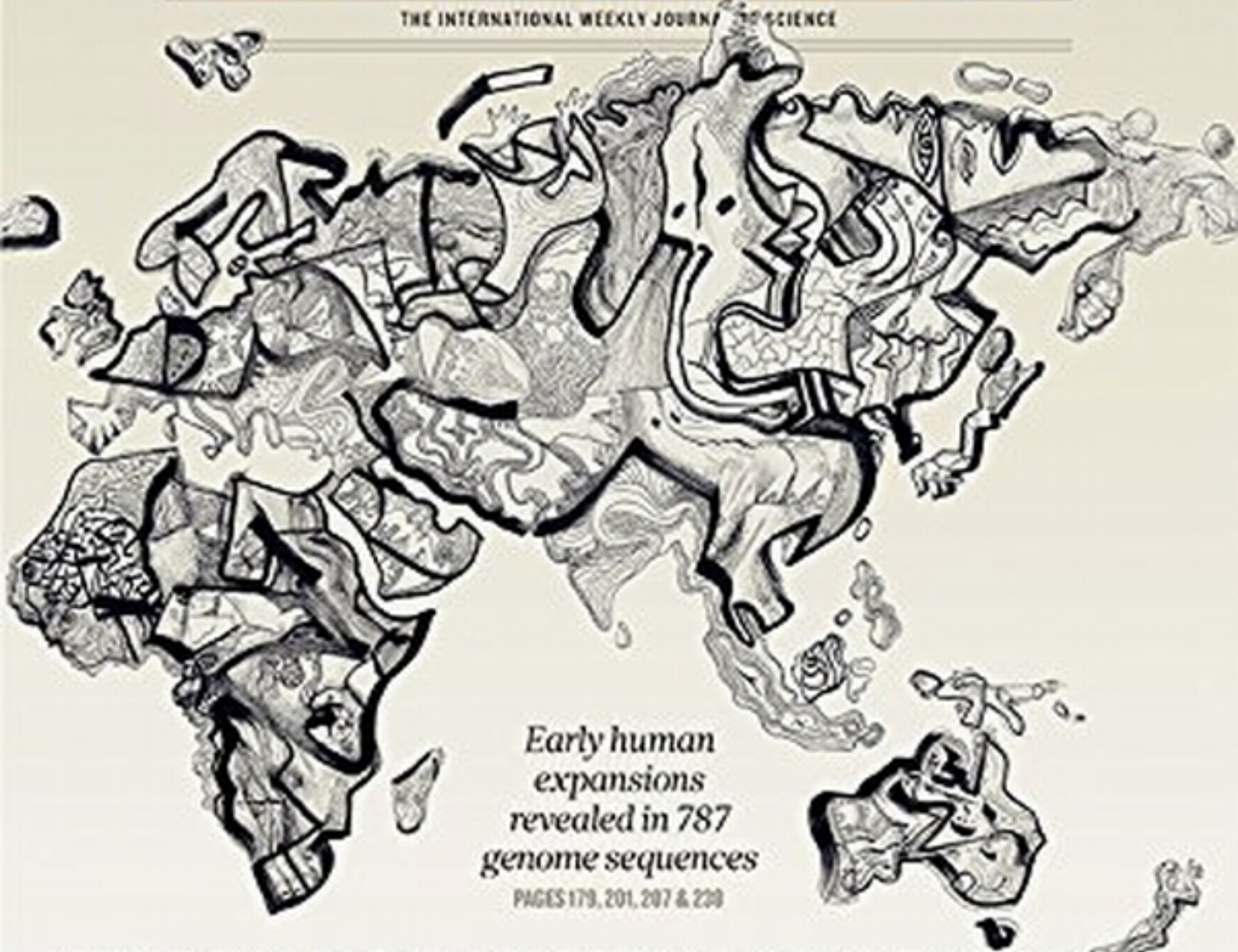


# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Early human  
expansions  
revealed in 787  
genome sequences  
PAGES 179, 201, 207 & 220

## THE DNA OF ANCIENT MIGRATIONS

### GENETIC DISEASES

#### DNA'S FALSE ALARMS

The 'lethal' mutations  
that pose no threat

PAGE 154

### PERSONALIZED MEDICINE

#### ALL GENOMES MATTER

Bias towards a 'European'  
genomics must be corrected

PAGE 161

### PLANETARY SCIENCE

#### CHANGING FACE OF THE MOON

NASA's orbiter measures the  
contemporary cratering rate

PAGES 177 & 215

NATUREASIA.COM

13 October 2018

Vol 558 No 7624

# THIS WEEK

## EDITORIALS

**BUZZWORDS** The nexus bandwagon heads in linguistic circles **p.140**

**WORLD VIEW** Data will not save the elephants or other species **p.141**



**EUREKA** Buoyancy aids discovered in floating diatoms **p.143**

## Healing traumatized minds

*Host countries need to deal with the raised levels of mental-health disorders in refugees if they expect them to integrate well, and that could mean benefits for psychological care in rich countries too.*

“**W**ir schaffen das!” proclaimed Germany’s chancellor Angela Merkel at the peak of the refugee crisis in August last year. “We will cope.”

Her promise met with widespread acclaim and the German public in turn proffered a *Willkommenskultur* or ‘culture of welcome’. But within months, as hundreds of thousands more refugees entered the country, the phrase began to hang heavy around Merkel’s neck — and could yet sink her. And the *Willkommenskultur* must confront a growing shock-wave of xenophobia, pushed ahead of the waves of displaced people.

Merkel’s phrase had of course been referring to the German population, who have to cope with finding accommodation, education and jobs for the refugees, as well as providing health services. This costly range of social adjustments will be more challenging even than the 1990 reunification. (Germany is the most popular destination in Europe for refugees.)

The refugees have to cope too. Yet in the highly charged public debate about the refugee crisis, the mental state of these vulnerable people rarely features — aside from vague and nakedly political warnings that some could be dangerous.

Refugees must cope with having been driven from their homes by violence and fear, and arriving in a foreign country with nothing. They need clear heads to make good decisions about their immediate and longer-term futures. They need flexibility of mind to adapt to their new, often disappointing, environments. They need to learn new things quickly, not least the language of their host countries, to meet the expectations that they will integrate quickly. But the existential stresses faced by the refugees at home and on their dangerous journeys has taken a disastrous toll on the minds of many of them.

### THERAPY DEVELOPMENT

In a News Feature on page 158 this week (of which Monday was World Mental Health Day), *Nature* examines some of the issues. The headline figure is deeply concerning. Psychologists in Germany estimate that more than half of those who have recently arrived there could be affected by post-traumatic stress disorder, depression, anxiety or another mental disorder. This is not a good basis for coping — for decision-making, adapting and learning. Such mental-health conditions reduce cognitive capacities and suck energy and motivation.

Medics and others who have worked with traumatized populations in far-flung war zones, such as Cambodia, Vietnam, sub-Saharan Africa, the Balkans and the Middle East, are familiar with this. The World Health Organization and the United Nations High Commissioner for Refugees published new guidelines last year that stress the importance of noticing and addressing mental-health issues. Yet host countries tend not to realize the extent of the problem, or to argue that mental-health problems can be tackled after the refugees are settled. But the right moment for support is during the volatile times. That makes sense for practical as well as humanitarian reasons — to ensure the smoothest

passage to integration for those who have no prospect of a return home.

Some German cities are rolling out some modest pilot programmes for psychiatric help, and Sweden will do so soon. But, considering the numbers affected, a very large investment is required across the continent. The European Union registered 1.4 million people seeking asylum in the 18 months up to June this year, and hundreds of thousands more may have entered without registering, according to

German estimates.

**“Refugees must cope with having been driven from their homes by violence and fear.”**

Access to mental-health provision is already difficult for many citizens. Prioritizing work with new arrivals is a tough sell. So it is important to consider the payback for the wider society.

Huge numbers of traumatized refugees in the Middle East and Africa are camped in countries with few psychiatrists, or are caught in areas too dangerous for aid workers to access. Their fate has accelerated efforts to develop simple and cheap therapies, some Internet- or app-based, as alternatives (or supplements) to conventional contact-heavy therapies. To broaden delivery, these can often be administered by trained lay people.

Clinical psychologists and psychiatrists now want to properly test these new therapies, which are based on the most up-to-date understanding of the brain and cognition, among refugees in Europe. Apart from the immediate relief they could bring, and the consequent chance of faster integration, there are two main reasons to encourage this.

These efforts will help to refine the therapies for application in all refugee centres, wherever and whenever war breaks out. And they will also help to break down barriers to modern approaches to clinical psychology in Europe, where the discipline has become conservative and complacent. Too many psychologists are reluctant to consider how mobile-device and Internet-based approaches could supplement standard therapies, and are too resistant to the concept that anyone who is not a qualified psychologist could help. The experience with refugees might also inspire improvements in local access to mental-health provision by generating, through necessity, a system that works faster and has fewer barriers.

Much could also be learnt from an ‘employment buddy system’ in Germany called ‘*Wir zusammen*’ or ‘We together’. The strongest signal of successful integration is entry into the work force, but this can be fraught with diverse and unpredictable problems. *Wir zusammen* is a movement of chief executives who create jobs or training positions for refugees that come with on-staff mentors — volunteers who oversee and champion their charges and who can accompany them to official appointments. This practical support could reduce the gnawing stresses that undermine mental health.

‘We together’ and ‘We will cope’ are feel-good phrases, but they should not be dismissed as platitudes. When it comes to mental health, they are fundamental. ■



# Genetic reckoning

*Researchers need to reassess many accepted links between mutations and disease.*

One of the major findings of the Exome Aggregation Consortium (ExAC), the largest-ever catalogue of genetic variation in the protein-coding regions of the human genome, is that many genetic mutations have been misclassified as harmful (M. Lek *et al. Nature* **536**, 285–291; 2016). Authors of that study estimate that each person has lurking in their genome an average of 54 mutations that are currently considered pathogenic — but that about 41 of these occur so frequently in the human population that they aren't in fact likely to cause severe disease. That finding is having major consequences for some people with such variants, lifting the equivalent of genetic death sentences (see page 154).

That raises two challenges for researchers: how to sort out which mutations currently considered pathogenic are actually benign, and how to apply more rigorous tests to future research that aims to find the genetic causes of disease.

Working out which mutations are actually linked to illness will be a long and arduous task. For instance, geneticist and physician Leslie Biesecker of the US National Human Genome Research Institute in Bethesda, Maryland, found that a patient referred to him for diagnosis harboured a genetic variant that had been linked to kidney failure. Yet it turned out that the variant was too common in ExAC to realistically be causing a rare kidney ailment. So Biesecker checked genome sequences from 950 people whom he had previously sequenced in a study called ClinSeq (K. L. Lewis *et al. PLoS ONE* **10**, e0132690; 2015). Five of them had the same variant, with no history of kidney disease, indicating that the variant probably does not actually cause this illness. To probe further, Biesecker is now recontacting the five ClinSeq participants with the variant to ask them to take part in follow-up tests to check whether they have normal kidney function, including collecting multiple urine samples over a 24-hour period.

To reassess the links between diseases and mutations, researchers must have access to a group of people whose detailed genetic and clinical information are known, and that's rare. It also takes time and some cost; multiply that by the huge numbers of 'pathogenic' variants that have been called into question, and researchers are looking at a major undertaking. It's a crucial one, because geneticists are being asked every day to make judgements about the harm that could

**“Many have not required enough evidence before asserting that a particular variant is harmful.”**

be caused by mutations found in patients' genomes. Biesecker hopes that planned or existing projects to link people's genomes to their detailed health records — such as the US president's Precision Medicine Initiative, which aims to sequence at least 1 million Americans, and the UK 100,000 Genomes Project — will help.

The rethink on pathogenicity shows that researchers who hunt for genetic mutations likely to cause disease need to be cautious. Many, it seems, have not required enough evidence before asserting that a particular variant is harmful.

Early efforts to discover the genetic underpinnings of disease started with families in which a particular condition recurred, generation after generation. By studying their extensive pedigrees, researchers could see strong evidence that certain mutations caused the disease. But in recent years, researchers have switched tactics: for instance, searching for evidence of pathogenicity by scanning for mutations that are more common in people with disease than in those without. It is becoming clear that many human genetic variations are relatively rare, and when researchers do not examine large enough groups of people with and without disease when scanning for pathogenic mutations, they are likely to mistakenly conclude that particular variants of interest turn up only in people with disease. The truth may be that they just haven't looked hard enough for these variants elsewhere.

These conclusions have consequences for real people, and so researchers must go about this work differently. When they suspect that a variant is linked to disease, they should check to see how common it is in databases such as ExAC. Even better, they should hunt for evidence that the mutation has a functional role in disease before declaring that it is pathogenic. Let the reckoning begin. ■

# Buzzword off

*‘Nexus’ is enjoying new-found popularity. But what does it actually mean?*

At *Nature* we like to think we have always been ahead of the curve, so it's pleasing to see that this journal was using the word 'nexus' some 140 years ago. We mentioned in a book review the “organic nexus between the motor and tactile centres” in the brain in November 1876, for example.

In the twenty-first century, the term nexus stands for more than its dictionary definition of a connection or focus point. It's a buzzword, especially when tagged on to the end of a string of associated nouns.

For example, an article in *Environmental Science and Policy* this month draws attention to the popularization of the phrase 'water–energy–food nexus' in debates over the use of natural resources (R. Cairns and A. Krzywoszynska *Environ. Sci. Policy* **64**, 164–170; 2016).

Language matters and, although Aaron Ellison argues in this week's World View on page 141 that the term “natural resources” itself should be retired, we'll skip that to examine Cairns and Krzywoszynska's main point: buzzwords are Orwellian and obfuscate even as they pretend to enlighten. Discussions of the nexus between connected crises, in other

words, can generate more linguistic heat than policy light.

The authors argue that “understandings and usage of the term nexus are plural, fragmented and ambiguous”. This is not always viewed as a negative by those who use the word, of course. When there is little of substance to say, it often helps to use language that acts as a mirror so finely polished that every reader can see their own agenda and interests reflected. There are echoes here of the way the term Anthropocene has been adopted and borrowed by a range of scientific disciplines, each of which wants a taste of the action.

“The term nexus appears to have something of a paradoxical quality,” say Cairns and Krzywoszynska, “being simultaneously unarguably true at a simple descriptive level, and yet confusingly unintelligible or meaningless to actors unfamiliar with the discourse.”

The motives behind the eagerness to jump on the nexus bandwagon are not always sinister. Honest brokers use the term to try to encapsulate that unspoken and undefined territory where the implications of one action bleed into another; when the equal and opposite reaction also has consequences.

But the risk is that containing this territory, however loosely, constrains it instead — and that the nexus becomes the focus of the analysis, rather than a natural consequence of studying the supporting problems.

Perhaps, like the most distant stars, the nexus is best viewed only with peripheral vision: we can see it's there, but we shouldn't focus our gaze directly on it lest its true nature slips from view. And, at the very least, it shows that we should choose our buzzwords with care. ■

CLARISSE M. HART/HARVARD FOREST



## It's time to get real about conservation

To protect endangered species from extinction, the ecological community must become more politically involved, argues **Aaron M. Ellison**.

**H**ow can scientists protect biodiversity? In the wake of August's Great Elephant Census, which revealed a precipitous decline in numbers throughout Africa, there were the usual calls from researchers for more and better data. Only if we know where and how many of each species there are, this argument goes, can we hope to conserve them. This is nonsense.

Better data will not save elephants, rhinos or any other species. An enormous number of individuals, academic institutions, local, state and national governments, and multinational and non-governmental organizations have been collecting, assimilating and organizing such data for decades, essentially fiddling while our biological heritage burns.

Of course, biodiversity data can be important for conservation, to suggest priorities and to draw attention to threatened and endangered species. But biodiversity data rarely drive conservation decision-making. Rather, in the vast majority of cases, they are used to bolster decisions made for other reasons. The decisions last week by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) to tighten trade in endangered species of sharks, parrots and pangolins shows this. Fascination, charisma and plush toys captured the imagination of the delegates, and journalism, political pressure and social-media campaigns pushed the decisions.

This week's Global Scientific Meeting in South Africa's Kruger National Park of the International Long-term Ecological Research network (ILTER) demonstrates the problem. With its long and enviable track record in integrating social dynamics into the study of ecological systems and engaging with policy- and decision-makers to develop conservation policies, ILTER is meeting in the lengthening shadows of faceless elephants, dehorned rhinos, vanishing gorillas and the many other threatened and endangered endemic species of a continent besieged by Al-Qaeda, Al-Shabaab, ISIS and Boko Haram. Many nations are embroiled in civil conflict and some are ruled by corrupt kleptocrats more interested in using the government purse to renovate their estates than to lift their populations out of poverty, much less to conserve their biodiversity or even adequately staff their 'paper parks'.

In this light, do we really need more scientific sessions on nitrogen cycling or drivers of biodiversity across scales? Sure, if the goal is simply to publish more abstruse papers and more data sets that will be read only by our friends and colleagues. But we should not delude ourselves that these sessions, or the data and scientific syntheses they yield, will help decision-makers find the energy and backbone to stop elephant poaching in Africa, clearcutting and burning in Indonesia, fracking and fouling of water supplies in North America or eating anything that walks with its back to the sky in China.

Rather, if biodiversity really matters for the planet, and is essential for humanity's well-being, we need to get real about what it will take to conserve it for future generations. I suggest three crucial actions that scientists can take, beginning right now.

First, stop referring to anything that isn't human as a 'natural resource'. Language matters, and this language suggests that the existence of other species is predicated on the benefits they provide for us. Natural historians and systematists have long asserted that we need to 'put names to faces' before we can care about non-human species. But even though we have already described and named millions of species, the precipitous decline of worldwide biodiversity makes it abundantly clear that naming species isn't enough.

Second, acknowledge that better data rarely lead to 'better' decisions (or at least to those decisions we think we would make if we were in charge). No amount of data can overcome visceral negative responses to bats, spiders or snakes, or positive ones to pandas, pangolins or baby seals. Decisions about which species to save — and which to triage to extinction — are based on raw emotion, the views of many different stakeholders and myriad political calculations. As the CITES process has demonstrated, data can be marshalled to support conservation decisions with broad-based support from a range of parties. But such consensus are increasingly hard to come by, the resulting CITES decisions still do not provide airtight protection, and as conflicts rage around the world and rapid economic growth continues to be prioritized over conservation in both developing and developed countries, biodiversity will continue to decline.

Third, more scientists must get actively involved in the political process. Calling, e-mailing and writing to political leaders is a small but necessary first step. Showing up for seemingly endless political meetings is a larger but necessary follow-up. If we're not in the room, our voices won't be heard. Volunteering for local, regional, national or international groups directly involved in conservation decisions is a bigger commitment. But if not us, who? And running for elected office would logically follow. If not now, when?

Scientists studying ozone depletion and climate change have shown that getting involved directly in the decision-making process can give scientists a place at the global table and a voice to help effect political change. Scientists who both study biodiversity and want to see other species persist and thrive must follow their example. ■

**Aaron M. Ellison** is the senior research fellow in ecology at Harvard University and a conservation commissioner in Royalston, Massachusetts.  
e-mail: [aellison@fas.harvard.edu](mailto:aellison@fas.harvard.edu)

**BETTER DATA  
WILL  
NOT SAVE  
ELEPHANTS,  
RHINOS OR  
ANY OTHER  
SPECIES.**



# RESEARCH HIGHLIGHTS

Selections from the  
scientific literature

## MICROBIOLOGY

### Protozoan protects the gut

Many single-celled microorganisms are harmful, but others regulate the immune responses of their animal hosts to guard against infections.

Some such organisms, called protozoa, live in the intestine, but have not been as well studied as their disease-causing counterparts. Miriam Merad at the Icahn School of Medicine at Mount Sinai in New York City and her colleagues identified a previously unknown protozoan, *Tritrichomonas musculus*, in the intestines of some laboratory mice. When this microbe colonized the guts of other mice, the animals exhibited an inflammatory response that protected against pathogenic *Salmonella* bacteria. However, the animals also showed increased susceptibility to inflammatory intestinal disease and colon tumours.

*Cell* 167, 444–456 (2016)

## QUANTUM COMMUNICATION

### Quantum secret kept for a day

A bit of information can be kept secure for 24 hours before being revealed — more than 5 million times longer than the previous record.

Quantum cryptography guards against eavesdroppers, but in secure voting and sealed-bid auctions, a message must remain unread and protected for a certain period of time. Routing the message through a pair of trusted ‘friends’ that are between the sender and receiver can delay and secure the message, but the friends would need to be located extremely far away from the sender and receiver to achieve a delay of more than

a few milliseconds.

Anthony Martin and his colleagues at the University of Geneva in Switzerland developed a protocol in which this kind of exchange happened 5 billion times, with encryption occurring at each round that built on that created previously. This allowed the authors to separate the sender and receiver computers from their ‘friends’ by just 7 kilometres while securing the bit for 24 hours.

*Phys. Rev. Lett.* 117, 140506 (2016)



RICK WILKING/REUTERS

## CLIMATE

### Megadroughts loom large

Climate warming looks set to plunge the American Southwest into decades-long drought by the end of the century.

Such ‘megadroughts’ have hit the region (Lake Powell on the Colorado River, pictured) during the past millennium. To calculate how changes in temperature, rainfall and soil moisture will affect the likelihood of such events, Toby Ault of Cornell University in Ithaca, New York, and his colleagues ran simulations using climate models and two

greenhouse-gas emission scenarios.

If emissions continue to rise unabated, the projected increase in regional mean temperature alone will boost the risk of a megadrought to 70–99% by 2100, depending on whether precipitation increases moderately, stays the same or decreases. If warming remains below 2°C compared to temperatures seen in the second half of the twentieth century, that risk falls to less than 66%.

*Sci. Adv.* 2, e1600873 (2016)

## COGNITION

### Human-like ape expectations

Chimpanzees and other great apes seem to understand the beliefs of others, suggesting that this ability is not unique to humans.

Researchers have long debated whether humans are the only primates to have a ‘theory of mind’ — the ability to attribute mental states such as desires and beliefs to others. To test this, Christopher Krupenye at Duke University

in Durham, North Carolina, Fumihiro Kano at Kyoto University in Japan and their colleagues monitored the gaze of chimpanzees, bonobos and orang-utans as they watched short videos. Two videos showed a person watching an object being hidden, and then searching for it. A third video tested the apes’ understanding of false beliefs by showing the object being moved while the person was not watching. As the person then prepared to search for the object, most apes looked in anticipation to the location where the person

falsely believed the object was hidden.

The authors suggest an implicit theory of mind predates human evolution. *Science* 354, 110–114 (2016)

## ECOLOGY

## Warmer forests store less carbon

Climate change might reduce the amount of carbon that forests can store, in part because photosynthesis decreases at high temperatures.

Emily Meineke and her colleagues at North Carolina State University in Raleigh studied willow oaks (*Quercus phellos*) in the local area, where — as in other large cities — pavement and other hard surfaces absorb and slowly radiate the Sun's heat. This increases the temperature in some urban regions to a level comparable to that predicted for the next century as climate warming continues. The team measured trees' photosynthetic rate, as well as factors such as water stress and pest prevalence, in hotter and cooler areas of Raleigh.

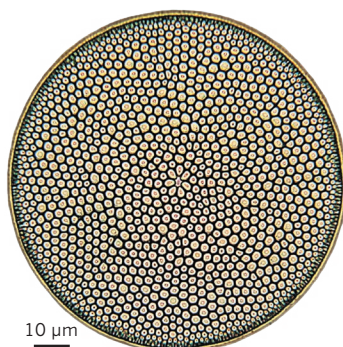
Hotter areas saw less tree growth. The team calculates that the 'urban heat island' effect reduced carbon sequestration in these trees by 12%. The reduction in growth was driven mainly by the effect of water deficits on photosynthesis, not increased herbivore activity. *Proc. R. Soc. B* 283, 20161574 (2016)

## MARINE BIOLOGY

## Diatoms sink in fits and starts

Single-celled marine organisms called diatoms can rapidly alter the speed at which they move through the water column, despite lacking structures for motility.

Diatoms are photosynthetic and are a major contributor to ocean productivity. Brad Gemmell and his colleagues at the University of Texas Marine Science Institute in Port



Aransas filmed three species of diatom (*Coscinodiscus radiatus* pictured) as they sank slowly in laboratory tanks. All exhibited stop-start movement: previously unobserved bursts of rapid sinking followed by periods of near-zero sinking. The authors visualized the flow of water around individual diatoms, and suggest that the organisms alter their buoyancy by exchanging ions with the seawater.

Rapid changes in sinking speed could explain how these diatoms compete for nutrients with cells that can actively swim.

*Proc. R. Soc. B* 283, 20161126 (2016)

## ASTRONOMY

## Strange fading star probed

A star seems to have been dimming for years, possibly because of a cloud of material obscuring it from view.

Benjamin Montet at the California Institute of Technology and Joshua Simon at the Carnegie Observatories, both in Pasadena, used instruments on NASA's Kepler spacecraft to study a star in the constellation Cygnus called KIC 8462852, which is brighter and larger than the Sun. Four years of observations revealed that the star dimmed slowly at first, by around 0.9% in total, then faded more rapidly by 2% in only six months. A few other stars nearby also became dimmer, but not to the same extent.

The authors speculate that the star's dimming could be explained by the collision or

break-up of a planet or comets in the star's system, creating a cloud of spreading debris.

*Astrophys. J. Lett. in the press*; preprint at <https://arxiv.org/abs/1608.01316> (2016)

## MICROBIOLOGY

## Gut bacteria help cancer drug

Certain gut microbes work with a common cancer drug by boosting anti-tumour immune responses, making the therapy more effective in mice.

Laurence Zitvogel of the Gustave Roussy Cancer Campus in Villejuif, France, and his colleagues studied the effect of two species of bacteria on the action of the drug cyclophosphamide. When they gave antibiotic-treated mice the microbe *Enterococcus hirae*, they found that it made immune cells called T cells more active against specific tumour markers and caused intestinal immune cells to proliferate. Another bacterium, *Barnesiella intestinihominis*, drove immune cells to infiltrate tumours. In mice that lack a protein that restricts these species' growth, the cancer drug was nearly twice as effective at reducing tumour size than in normal animals.

The results suggest that gut bacteria could be used to optimize cancer therapies, the authors say.

*Immunity* <http://doi.org/brmm> (2016)

## MATERIALS

## Supercapacitor made from MOF

Researchers have built a high-capacity energy-storage device using a metal–organic framework (MOF) — a porous material with many desirable properties.

MOFs are networks of metal ions linked together by organic molecules, and their large surface area means they hold promise as energy-storing supercapacitors. Mircea Dincă of the Massachusetts Institute of Technology in Cambridge

and his colleagues created such a device with electrodes made only from a nickel–organic framework ( $\text{Ni}_3(\text{HITP})_2$ ) that has high electrical conductivity. The supercapacitor stored more energy per area than most other carbon-based devices, and retained more than 90% capacity after 10,000 cycles — on a par with commercial devices.

Such a supercapacitor could have an important role in future energy grids, the authors say.

*Nature Mater.* <http://dx.doi.org/10.1038/nmat4766> (2016)

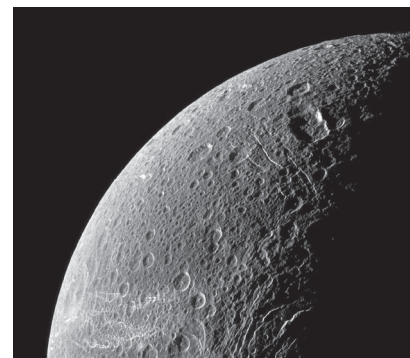
## PLANETARY SCIENCE

## Ocean on another of Saturn's moons

Like its neighbours Titan and Enceladus, Saturn's moon Dione may harbour an ocean beneath its icy surface.

Mikael Beuthe and his colleagues at the Royal Observatory of Belgium in Brussels studied data collected from Enceladus and Dione by NASA's Cassini spacecraft. They looked for small changes in the moons' gravity and shape that can reveal layers of buried liquid. Data modelling suggested that Dione has a 65-kilometre-deep global ocean hidden beneath some 100 kilometres of ice.

Those waters are a possible habitat for extraterrestrial microbes, should they exist. *Geophys. Res. Lett.* <http://doi.org/brg7> (2016)



➔ **NATURE.COM**

For the latest research published by Nature visit:

[www.nature.com/latestresearch](http://www.nature.com/latestresearch)



# SEVEN DAYS

The news in brief

## FACILITIES

### LIGO in India

A planned Indian outpost of the Laser Interferometer Gravitational-Wave Observatory (LIGO) will probably be built in Marathwada in the west of the country, after the state of Maharashtra announced on 4 October that it had approved a 40-hectare site for the laboratory. LIGO-India will join a global network of gravitational-wave sensors; these include the two US sites, in Washington state and Louisiana, that this year reported the first detection of the waves.

### Ukraine joins CERN

Ukraine has become an associate member of CERN, Europe's particle-physics lab near Geneva, Switzerland. The move allows Ukraine to hold and attend meetings of the CERN Council, but without voting rights. Yuriy Klymenko, Ukraine's ambassador to the United Nations in Geneva, called the move, announced on 5 October, "an extremely important step on the way [to] Ukraine's European integration". Ukraine becomes CERN's fifth associate member, alongside Serbia, Cyprus, Turkey and Pakistan.

## EVENTS

### Foreign experts

Academics were enraged last week by the assertion that the UK government had barred foreign academics from advising on Brexit projects. Reports emerged on 7 October that the Foreign and Commonwealth Office (FCO) had told the London School of Economics and Political Science (LSE) that it would not take advice from academics who did not hold UK passports. Whether this

communication did actually happen is unclear. The LSE maintains that it did, but the FCO — without commenting directly on the LSE's claim — has said that it will continue to take advice from "the best and brightest minds, regardless of nationality".

### Detained physicist

On 7 October, France's supreme administrative court, the State Council, refused an appeal by particle physicist Adlène Hicheur to have his house arrest lifted. Hicheur has previously been jailed in France for terrorism-related offences — charges vigorously disputed by his colleagues — and had restarted his research life in Brazil following his release in 2012. In July this year, he was

deported to France and placed under house arrest, under state-of-emergency powers introduced there following terror attacks. His ordeal has been widely condemned by his physicist colleagues. In a further twist last weekend, Hicheur, who has French and Algerian nationalities, sought to extricate himself from the situation by requesting that his French nationality be revoked and that he be immediately expelled from France.

### Kennewick man

The US House of Representatives has passed legislation to return controversial human remains to Native Americans in Washington state. The fate of the 8,500-year-old 'Kennewick

Man', a near-complete human skeleton, has been debated since its discovery in 1996. The case pitted scientists who wished to study the remains against Native American tribes that saw them as belonging to an ancestor. The bill must be reconciled with similar legislation passed last month by the Senate, and it will require the president's signature.

## CLIMATE CHANGE

### Aviation emissions

After 3 years of debate, negotiators from 191 countries struck a deal to reduce carbon emissions from international aviation. The agreement, made on 6 October at a meeting of the International Civil Aviation Organization in



CARLOS GARCIA RAWLINS/REUTERS

## Haiti faces cholera in hurricane's wake

Hurricane Matthew has swept a path of destruction across the Caribbean, killing at least 1,000 people when it hit Haiti on 4 October. Fears of a cholera outbreak are rising there in the wake of the storm, with at least 13 deaths from the disease reported so far. Matthew was the first category-5 hurricane in the Atlantic since Hurricane Felix in 2007. Its high winds

and heavy rains continued north but skirted the coast of Florida, where damage was less than originally feared. Only minor damage was reported by NASA's Kennedy Space Center, where the next-generation GOES-R weather satellite is awaiting launch to improve forecasts of, among other things, Atlantic hurricanes. The US death toll was reported to be at least 20.

JEFF CHIU/AP PHOTO

Montreal, Canada, establishes standards for energy efficiency and an emissions-offsetting programme that will see the aviation industry invest in projects to reduce emissions in other sectors. The programme could cover an estimated 75% of emissions growth from 2021–35, resulting in 2.5 billion tonnes of carbon offsets — equivalent to more than 6 years of UK carbon emissions. Aviation accounts for 2% of global carbon emissions.

## Climate amendment

United Nations delegates meeting in Kigali, Rwanda, are finalizing an amendment to the Montreal Protocol whereby the treaty will target greenhouse gases as well as substances that destroy ozone. The decision to rapidly phase out hydrofluorocarbons — refrigerants that are powerful greenhouse gases — was taken last year, and described as “welcome — and long overdue” by *Nature* at the time (see *Nature* 527, 133; 2015). Delegates are likely to take until the meeting’s final day on 14 October to hammer out the details of the amendment.

## BUSINESS

### Theranos cuts

Beleaguered blood-testing company Theranos of Palo Alto, California, announced in an open letter on 5 October that



it is closing its clinical labs and laying off some 340 employees — about 40% of its staff. The company had faced huge scepticism over its claims that it could perform hundreds of diagnostic tests using a single drop of blood. In July, US regulators banned its chief executive, Elizabeth Holmes (pictured), from running a lab for two years. Holmes said that the company will continue to develop its miniaturized blood-testing device, miniLab. Further troubles emerged on 10 October, when a shareholder company announced that it would be suing Theranos to recoup its investment of nearly US\$100 million, which it says was secured on the basis of lies.

## RESEARCH

### Telescope time bias

Female scientists are allocated less telescope time at the European Southern Observatory than are their male colleagues, according to an analysis posted on

arXiv on 5 October (F. Patat Preprint at <https://arxiv.org/abs/1610.00920>; 2016). The study of more than 13,000 proposals and 3,000 principal investigators found that 16% of proposals submitted by women were successful, whereas men had a success rate of 22%. Much of the disparity seems to be because men submitting proposals have more senior positions on average than women, the study says, and astronomers at higher career levels received greater ratings for their proposals during the review process.

## AWARDS

### Chemistry Nobel

Jean-Pierre Sauvage, Fraser Stoddart and Bernard Feringa shared the 2016 Nobel Prize in Chemistry, announced on 5 October, for their work on creating tiny molecular machines. The three have made molecular knots, shuttles, rotors, chains, pumps, axles, switches, memory devices and even a nanocar — all at the molecular scale. The nanoscale machines are yet to find applications, but researchers hope that their uses could range from delivering drugs to computer memory (see page 152). Separately, Yoshinori Ohsumi, the cell biologist who won the 2016 medicine Nobel, announced that he would use

## COMING UP

### 16–21 OCTOBER

The American Astronomical Society’s planetary-sciences division and the European Planetary Science Congress meet in Pasadena, California. [aas.org/meetings/dps48](http://aas.org/meetings/dps48)

### 1–10 NOVEMBER

The first Berlin Science Week brings together academics and institutions from all over the world to discuss science and society. [www.berlinscienceweek.com](http://www.berlinscienceweek.com)

the 8 million Swedish kronor (US\$940,000) awarded to establish a system to provide support for young researchers over decades.

## Research jackpot

A US\$75-million grant for research into coronary heart disease has been awarded to a team led by Calum MacRae, chief of cardiovascular medicine at Brigham and Women’s Hospital in Boston, Massachusetts. The money is from the One Brave Idea fund, set up by the American Heart Association, Verily Life Sciences (owned by Google’s parent company, Alphabet) and drug firm AstraZeneca to fund research on the most common type of heart disease in the United States. The grant, an unusually large sum for a single team, will be delivered in \$15-million chunks over five years.

## CORRECTION

The item ‘Nuclear go-ahead’ (*Nature* 537, 455; 2016) should have said that Hinkley Point C is expected to supply 7% of UK electricity — not energy — demand.

## TREND WATCH

Investors fled on 5 October when a drug company focused on RNA-interference (RNAi) therapies announced that it would abandon one of its lead drug candidates amid safety concerns. The drug, revusiran, was in phase III clinical trials to treat a form of amyloidosis. The news sent stock in Alnylam Pharmaceuticals of Cambridge, Massachusetts, plummeting by roughly 50%. Companies have struggled to translate RNAi into therapies, and investors have been lukewarm on biotech stocks this year.

### RNAi-DRUG FAILURE SENDS STOCK TUMBLING

Alnylam Pharmaceuticals’ share price crashed by about 50% on 5 October, after the company halted a phase III trial of a drug that uses RNA interference.



SOURCE: GOOGLE FINANCE



# NEWS IN FOCUS

**PLANETARY SCIENCE** NASA rethinks approach to Mars exploration **p.149**

**CANCER RESEARCH** Promising therapy raises safety concerns **p.150**

**CHEMISTRY** Molecular machines nab Nobel prize **p.152**



**PSYCHIATRY** Rise in mental-health disorders in Europe's migrants and refugees **p.158**

CHRIS RATCLIFFE/BLOOMBERG/GETTY



UK Prime Minister Theresa May gives the closing speech at the Conservative Party conference, where she took a hard line on immigration.

## POLITICS

# Scientists spooked by UK anti-immigration stance

*Plans to restrict freedom of movement intensify fears over June's Brexit vote.*

BY DANIEL CRESSEY

UK scientists say they're dismayed by their new government's toughened stance on curbing immigration, including ideas to restrict the flow of foreign students and workers.

The government outlined its plans last week at the annual Conservative Party conference, which was the first since Theresa May became prime minister in the wake of June's vote for Brexit — the decision that the United Kingdom should leave the European Union. At the

conference in Birmingham, politicians made it clear that they want to eliminate the free movement of EU citizens into the United Kingdom once the country splits from the EU, an event now expected to take place in 2019. "We are not leaving the European Union only to give up control of immigration again," said May.

Since the referendum, in which concerns over immigration were believed to have played a big part in swaying voters, scientists have worried about how Brexit would affect the free movement of people. Although the government has not fleshed out its latest proposals,

they are the strongest indication yet that scientists will not be allowed to move freely between the United Kingdom and the EU after Brexit — which in turn means that UK researchers may be excluded from EU funding programmes.

"There has been a change in tone. I was surprised by how strong some of the comments were," says Azeem Majeed, who heads the department of primary care and public health at Imperial College London. The perception that non-UK citizens are not welcome — which grew as a result of June's Brexit referendum — has only increased since the conference, ▶

► he says. That is particularly the case in health fields, because the government has pledged to cut the number of jobs for given to foreign doctors, in favour of UK citizens.

UK universities get about 16% of their research funding and 15% of their staff from the EU, and scientists have been vocal about the need to maintain freedom of movement. It may even be a prerequisite for UK access to EU research funding. When Switzerland restricted freedom of movement in 2014, its researchers lost access to the major Horizon 2020 research-funding programme, leading to protracted negotiations that are still ongoing. “The hard line on freedom of movement is almost certain to restrict us from EU funds,” says Stephen Curry, a structural biologist at Imperial College London. Other comments on immigration and restricting foreign students are also going down poorly in academia, he says. “It’s reinforcing the rather sour atmosphere.”

#### WORKER RESTRICTIONS

UK home secretary Amber Rudd said in her conference speech that the government would consider making it harder to recruit from overseas, forcing companies to publicly disclose the proportion of foreign staff working for them (an idea that other politicians later disavowed) and cutting down on universities’ ability to recruit foreign students to “lower quality courses”.

“We had a very decisive message from the Conservative conference that the priority is simply reducing the number of people who come here, and if that damages the economy, so be it,” says Jonathan Portes, an economist at the UK National Institute of Economic and Social Research in London.

However, the conference was not all bad news for science and science policy, says Sarah Main, director of the Campaign for Science and Engineering in London. She says that the comments on immigration are concerning. But she adds that, at the Birmingham conference, “We’ve seen the government being much more clearly positive about research and innovation in general.”

She cites a speech from Chancellor of the Exchequer Philip Hammond — which praised science as a driver of growth and emphasized the need to get “the brightest and best to work here in our high-tech industries” — and positive comments from science minister Jo Johnson at events away from the main auditorium.

But Portes, chief economist for the UK Cabinet Office during the 2008–09 financial crisis, says that Hammond’s positive messages for science don’t outweigh the negative impacts of May and Rudd’s plans. “It’s nice to know the chancellor is not on the same page as the PM and the home secretary, but it seems pretty clear who is calling the shots,” he says. ■



Schoolchildren in Limpopo, the site of one of South Africa’s existing demographic surveillance studies.

#### POPULATION STUDIES

# South Africa plans huge health study

*Network would be Africa’s largest demographics project if it can sustain long-term funding.*

BY LINDA NORDLING

South Africa’s government has announced that it will expand the country’s existing demographic studies to create the largest project of its kind in Africa — tracking the health, income and educational attainment of around 1% of South Africa’s population.

The Department of Science and Technology estimates that it will put 264 million rand (US\$19 million) into the demographic project over the next five years, which will eventually cover at least half a million people. It has secured the funding for its first three years; the rest will need to be allocated in future government budgets.

If the study can be sustained in the long term, researchers hope that the data will help them to track efforts to curb major health problems such as HIV and tuberculosis, and to monitor emerging lifestyle-related threats such as cancer and diabetes. The department

intends the survey to run for decades, following people from the cradle to the grave and monitoring intergenerational trends.

Long-term demographic studies have played an important part in charting disease patterns. One survey that began in 1948 in Framingham, Massachusetts, led to the discovery of cardiovascular disease risk factors such as smoking and diabetes, and has allowed scientists to study intergenerational disease patterns.

But in Africa, as in many other parts of the developing world, such long-term projects have been neglected in favour of a focus on health emergencies such as HIV or Ebola, says Glenda Gray, president of the South African Medical Research Council.

“You never get your head above water to plan for the future,” she says. She thinks that the new project will change that.

South Africa has had three demographic surveillance projects running since the mid-to late 1990s, based in Mpumalanga and

NDLOVU/SOWETAN/GALLO IMAGES/GETTY



Limpopo in the northeast, and KwaZulu-Natal on the east coast. These have been able to track trends such as a growth in life expectancy as the country rolled out antiretroviral drugs to fight the HIV epidemic. But the long-term sustainability of such studies — which have been funded by non-governmental donors — is a perennial concern, says Kobus Herbst, deputy director of the African Health Research Institute, based in Durban, which runs the study in KwaZulu-Natal. So the government's investment is particularly welcome, he says.

#### FROM RURAL TO URBAN

All the existing surveillance projects are in rural areas, providing only a narrow view of national population trends, says Gray. “The rural sites have been critical for understanding things like how antiretroviral rollout plays out in districts,” she says. But they don't catch emerging patterns of disease linked with modern city life, driven by factors such as pollution, work-related stress and dietary changes. Of the four new surveillance nodes in the planned network, three will be based in South Africa's biggest cities: Cape Town, Johannesburg and Durban (see ‘Tracking South African health’).

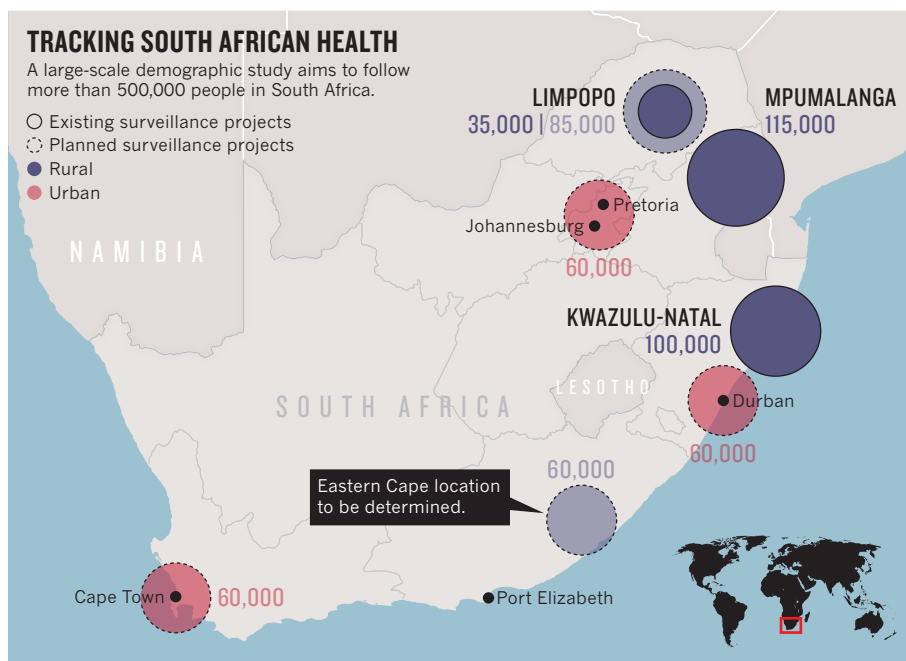
The existing surveys already cover around 250,000 people, but each collects different types of data, so their measurements cannot be compared or integrated together. During the first three years, the surveys will be linked up, and the Limpopo one will be expanded. By the end of the three-year period, a total of 300,000 people should be included in the project, says Herbst. The target of 500,000 people will be reached, hopefully, in five years' time, he says.

The government funding will cover the full

#### TRACKING SOUTH AFRICAN HEALTH

A large-scale demographic study aims to follow more than 500,000 people in South Africa.

- Existing surveillance projects
- Planned surveillance projects
- Rural
- Urban



health and socio-demographic surveys, and will fund linkages to national health records and the collection of dried blood spots from adult participants once a year for HIV testing, Herbst says. To do more — such as DNA sequencing — will require funding from external donors.

Linda Fried, an epidemiologist who is dean of Columbia University's Mailman School of Public Health in New York City, thinks that the surveys will not only allow South Africa to develop its science base but will also attract international investment.

The programme was launched on 4 October

by South African science minister Naledi Pandor, at a conference to plan out South Africa's first road map for national research infrastructures. In addition to the demographic project, the road map launched this week includes plans for a nuclear-medicine research facility dedicated to drug development and clinical research, a solar-research facility to demonstrate photovoltaic technologies, and a new hub to coordinate efforts to protect the country's natural-history collections.

“We build big scientific infrastructure to attract international researchers to our country,” Pandor said. ■

#### PLANETARY SCIENCE

# NASA rethinks Mars exploration

*Agency considers time-allocation model in an era of shifting international interests.*

BY ALEXANDRA WITZE

NASA is investigating a new way of studying Mars. Starting in the 2020s, scientists who participate in the agency's Mars missions might no longer design and build their own highly specialized payloads to explore the red planet. Instead, planetary scientists could find themselves operating much as astronomers who use large telescopes do now: applying for time to use a spacecraft built with a generic suite of scientific instruments.

The proposed change is spurred by NASA's waning influence at Mars. The agency's

long-running string of spacecraft is winding to a close, and international and commercial interests are on the rise. By the middle of the next decade, European, Chinese, Emirati and SpaceX missions are as likely to be at Mars as NASA is (see ‘Red-hot planet’).

Jim Watzin, head of NASA's Mars exploration programme in Washington DC, suggested the new approach to the red planet on 6 October at a virtual meeting of Mars scientists. “The era that we all know and love and embrace is really coming to an end,” he said. “It's important to recognize that the future is not going to be the same as the past.”

Throughout the 2000s, NASA sent a

sustained barrage of spacecraft to Mars, unique in the sheer number of robots directed at one planetary target. But many have expired, and the ones still operating are growing old. NASA's three functional orbiters — Mars Odyssey, Mars Reconnaissance Orbiter, and MAVEN — launched in 2001, 2005 and 2013 respectively. The Opportunity rover is in its thirteenth year, and the Curiosity rover is in its fifth.

NASA has only one more spacecraft scheduled in its Mars programme, a rover due to launch in 2020 that is tasked with gathering samples for an as-yet-unscheduled return to Earth. (The InSight geophysics probe, slated for a 2018 launch, was not developed under ▶

## RED-HOT PLANET

The United States still has a fleet of spacecraft exploring Mars, but other countries — and commercial interests — are joining the effort.

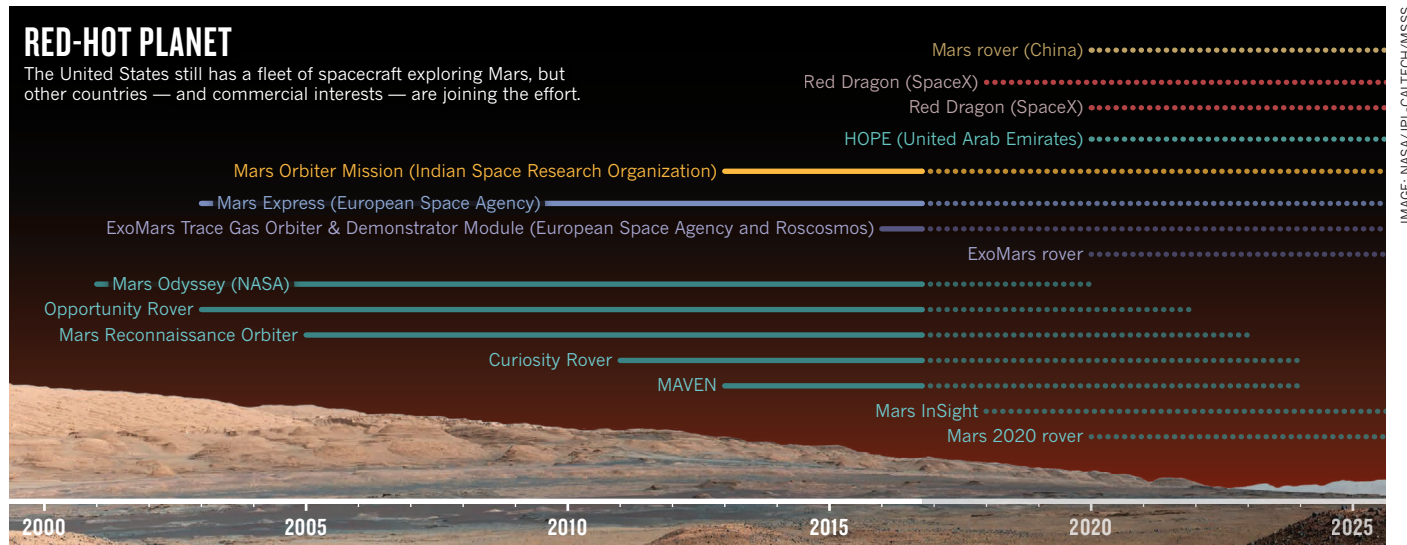


IMAGE: NASA/JPL-CALTECH/MSS

► the auspices of NASA's Mars programme.)

NASA wants to start planning for an orbiting mission to launch after 2020. In June, the agency asked five companies about what sorts of Mars orbiters they might be able to build, and how quickly and cheaply that could be done. Five international partners have also said they would like to be involved, Watzin said.

Many non-NASA missions to Mars are already on the books. In 2020, the European Space Agency and China each plan to launch Mars rovers, while the United Arab Emirates will send an orbiter. SpaceX of Hawthorne, California, hopes to send its first Red Dragon landers to Mars in 2018.

This broadening context prompted Watzin to propose the new way of operating Mars

missions. "I'm not trying to fix something that's broken," he said. "I'm trying to open the door to a larger level of collaboration and participation than we have today."

In the facility-based approach, scientists would propose investigations using one or more instruments on a future spacecraft. NASA would award observing time to specific proposals, much as telescope-allocation committees parcel out time on their mountaintops. This would be different from the current approach, in which individual teams of scientists propose, build and operate instruments.

Watzin's proposal is a trial balloon, not an official change to NASA policy. "It's a little early yet to figure out how the community is going to respond," says Jeffrey Johnson, a

planetary scientist at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland, and head of the group that organized the meeting.

But some researchers are already pushing back. Alfred McEwen, a planetary scientist at the University of Arizona in Tucson, noted that the Mars Reconnaissance Orbiter's HiRISE camera has taken thousands of images of Mars based on public requests. "We've managed to do all the things [Watzin] described already without a new paradigm," says McEwen, the camera's principal investigator. "We have distributed operations, we have multiple customers, we have a foreign contributed instrument. So my immediate reaction to this idea was not very positive." ■

## DRUG DEVELOPMENT

# Safety concerns blight promising cancer therapy

*As the first T-cell treatments for tumours near US approval, researchers race to engineer less-toxic versions.*

BY HEIDI LEDFORD

A groundbreaking treatment that arms immune cells called T cells to battle cancer is barreling towards regulators, fuelled by unprecedented clinical success and investor exuberance.

But progress of the therapy, called CAR-T, has been marred by its toxicity; several deaths have been reported in clinical

trials. Even as the first company readies its application to the US Food and Drug Administration (FDA) — expected by the end of the year — researchers are hard at work to make the supercharged T cells safer.

Doing so is crucial to expanding the use of the therapy to more people, says Anthony Walker, a managing partner at Alacrita, a consulting firm in London. "Right now it is heroic medicine," he says — a gruelling

treatment deployed only in people for whom all else has failed. "Patients are taken sometimes to within an inch of their lives."

Most CAR-T procedures begin by harvesting a patient's white blood cells and sifting out the T cells. Those T cells are engineered to recognize cancer cells, and then infused into the patient, ready to do battle. The approach has shown remarkable success against leukaemias and lymphomas: in one



study, all traces of leukaemia disappeared in 90% of the patients who received the treatment (S. L. Maude *et al.* *N. Engl. J. Med.* **371**, 1507–1517; 2014).

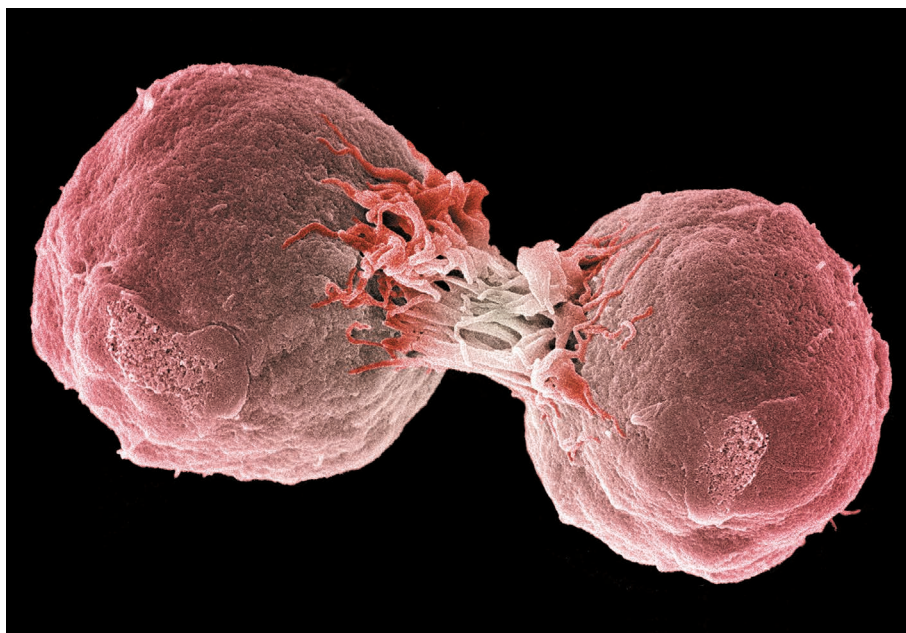
Results such as those have fuelled an investor frenzy. “It set the field on fire,” says Walker. Swiss pharmaceutical giant Novartis invested in the technique in 2012. In 2014, CAR-T firm Kite Pharma of Santa Monica, California, raised US\$128 million when it went public. A few months later, one of its competitors, Juno Therapeutics of Seattle, Washington, yielded \$264 million in its initial public offering.

Now Kite is racing to be the first to bring a CAR-T therapy to the market. On 18 October, the company will update investors on its plans to manufacture and sell the complex therapy, which it hopes to launch in 2017.

But the treatment’s toxicity has discouraged some investors. On 26 September, Kite released interim clinical-trial results — widely seen as successful — in people with aggressive non-Hodgkin’s lymphoma (see [go.nature.com/2djdqen](http://go.nature.com/2djdqen)). Yet about one-third of the patients developed serious neurological side effects, and 18% developed a deadly condition called cytokine release syndrome, which can cause organ failure. Two of the 62 patients died as a result of the treatment.

That toxicity is unlikely to dissuade the FDA, given the dramatic effects of the treatment, says analyst Michael Yee of the investment bank RBC Capital Markets in San Francisco, California. “It has transformed what was essentially a death sentence into a potential for long-lasting remission,” he says.

But the toxicity does leave room for improvement. One approach that researchers are studying to boost safety is to improve standardization of each patient’s dose of T cells. CAR-T therapies typically begin with a mixture of various kinds of T cell, some with very different functions. “Not all T cells are created equal,” says Stanley Riddell, an immunologist at the Fred Hutchinson Cancer Research Center in Seattle. To create a better-defined T-cell cocktail, Riddell’s lab first sorts out different types of T cell and blends them together again in specific proportions. So far, he says, trials in 140 patients suggest that the approach provides better control of dosage — and toxicity (see, for example, C. J. Turtle *et al.* *J. Clin. Invest.* **126**, 2123–2138; 2016).



Dividing lymphoma cells, which can be destroyed by CAR-T therapy.

Other groups have developed a ‘suicide switch’ to shut off the CAR-T cells in the body. If toxicity is spiralling out of control, doctors can administer a drug that activates the switch — a modified version of a protein called caspase-9 — and triggers the CAR-T cells to self-destruct.

That approach has not been popular with clinical researchers, who often opt to treat

**“It has transformed what was essentially a death sentence into a potential for long-lasting remission.”**

toxic reactions with other drugs rather than risk shutting down the treatment altogether, notes Michel Sadelain, an immunologist at Memorial Sloan Kettering Cancer Center in New York City. “What you’re doing is destroying your extremely expensive medication after you’ve administered it,” says Walker.

But cancer researcher Malcolm Brenner of Baylor College of Medicine in Houston, Texas, says that the switch is not all or nothing: adding just a little of the activation drug can dampen toxic effects without killing all of the engineered T cells.

Michael Brown, a clinician and cancer

researcher at the Royal Adelaide Hospital in Australia, is using the suicide-switch approach in his CAR-T clinical trial against melanoma. So far, he says, his patients haven’t needed it. But having the switch in place helped him to feel more secure about the trial, in which T cells target a protein that is more abundant in melanoma but is also expressed at low levels in normal brain tissue.

Many researchers hope to reproduce CAR-T’s success against leukaemia in solid tumours such as melanoma. And, like Brown, they are struggling to find proteins on cancer cells that could serve as targets for T cells but that are absent from normal tissues. One way could be to focus on multiple proteins expressed by cancer cells, says Sadelain. The therapy will then attack only cells that express all of those proteins, to provide a more precise way to mark tumour cells for destruction.

For now, all eyes are on Kite to see whether it can get its therapy approved by regulators and into hospitals. Even if the company succeeds, Walker is betting that the treatment will be the first in a long line of CAR-T therapies. “We’re at such an early stage in this field,” he says. “If you wind the clock forward 10–15 years, I think it will be unrecognizable.” ■

  
**MORE  
ONLINE**

#### TOP NEWS



Human age limit claim sparks debate  
[go.nature.com/2e5QTuv](http://go.nature.com/2e5QTuv)

#### MORE NEWS

- Pangolins and parrots among winners at largest-ever meeting on wildlife trade [go.nature.com/2du0ehf](http://go.nature.com/2du0ehf)
- Apes can tell when you’ve been duped [go.nature.com/2dlkcy](http://go.nature.com/2dlkcy)
- World leaders discuss refrigerant ban [go.nature.com/2e2wuun](http://go.nature.com/2e2wuun)

#### NATURE PODCAST



Refugee mental health; better neural nets; and changing attitudes to female genital cutting [nature.com/nature/podcast](http://nature.com/nature/podcast)

## EDUCATION

## Where Nobel winners start

*List reveals undergraduate schools with most laureates.*

BY TOM CLYNES

A new study ranks institutions by the proportion of their undergraduates that go on to win Nobel prizes.

Two schools dominate: École Normale Supérieure (ENS) in Paris and the California Institute of Technology (Caltech) in Pasadena. These small, elite institutions each admit fewer than 250 undergraduates per year, yet their per capita production of Nobelists outstrips some larger world-class universities by factors of hundreds.

"This is a way to identify colleges that have a history of producing major impact," says Jonathan Wai, a psychologist at Duke University in Durham, North Carolina, and a co-author of the unpublished study.

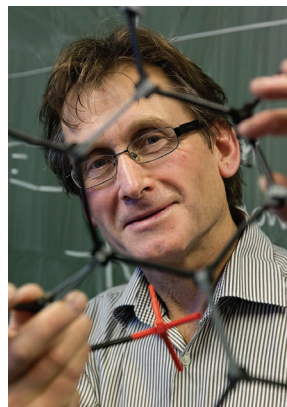
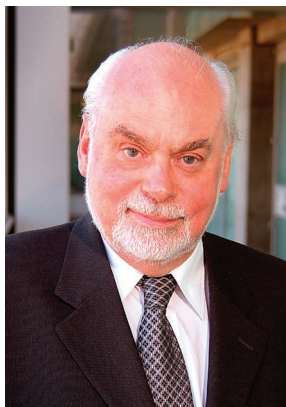
Wai and Stephen Hsu, a physicist at Michigan State University in East Lansing, examined the 81 institutions worldwide that each had at least three alumni who received a Nobel award in any of the six categories between 1901 and 2015. For a meaningful comparison, the team divided the number of Nobel laureates at a university by its estimated number of undergraduate alumni.

Many of the top Nobel-producing institutions are private, with significant financial resources. Among the more surprising high performers were several small US liberal-arts colleges, such as Swarthmore College in Pennsylvania (ranked at number 4) and Amherst College in Massachusetts (number 9).

To gauge trends over time, Wai cut the sample of 870 laureates into 20-year bands. US universities, which now make up almost half of the top 50 list, began to dominate after the Second World War. Whereas French representation in the Nobel ranks has declined over time, top-ranked ENS has remained steady in its output.

Santo Fortunato, a theoretical physicist at Indiana University Bloomington who has researched trends in Nobel prizewinners, cautions that the methodology cannot produce a highly accurate or predictive ranking. "There is a high margin of error due to the low numbers of prominent scholars," he says.

Wai and Hsu agree that there are statistical uncertainties in their rankings, owing to the small number of prizes awarded each year. The two are confident that the ENS and Caltech lead the pack, but say that statistical fluctuations could change the order of institutions placed from third to ninth, Hsu says. ■



Molecular architects: Fraser Stoddart, Bernard Feringa and Jean-Pierre Sauvage.

## NANOTECHNOLOGY

## Chemistry Nobel for nanomachines

*Award recognizes three pioneers of molecular motors.*

BY RICHARD VAN NOORDEN & DAVIDE CASTELVECCHI

Three chemists who created tiny molecular machines have won the 2016 Nobel Prize in Chemistry for their intricate designs.

Jean-Pierre Sauvage, at the University of Strasbourg in France; Fraser Stoddart, a Scottish-born chemist at Northwestern University in Evanston, Illinois; and Bernard Feringa, at the University of Groningen in the Netherlands, share the award for their work in the 1980s and 1990s, when they pioneered efforts to miniaturize motors.

"I'm a bit shocked because it was such a great surprise. And I'm so honoured," said Feringa in a telephone interview with the Nobel Committee just after the prize-winners were announced in Stockholm on 5 October.

The three have made molecular knots, shuttles, rotors, chains, pumps, axles, switches, memory devices and even a nanocar — all at the scale of molecules (see 'Nanomachines'). The nanoscale machines are yet to find application, but researchers hope that their uses could range from delivering drugs to computer memory.

"It's early days, of course," Feringa told the Nobel Committee. "But once you are able to control movement, you have a motor, you can think of all kinds of functions." He suggested that the machines could be used as tiny robots in the body to deliver drugs or detect cancerous cells; or as smart materials

that could adapt or change depending on external signals.

"I applaud the fact that — for once in chemistry — Stockholm has recognized a piece of chemistry that is fundamental in its making and being," Stoddart said at a press conference at Northwestern University, held later in the day.

Only a handful of laboratories are currently actively engaged in making nanomachines, says Dean Astumian, who studies the theory of molecular motors at the University of Maine in Orono. But he thinks the field will get a boost from the award. "The recognition that is afforded by a Nobel prize is going to attract the best young people," he says. Astumian thinks the work will provide applications within 25 years. "There's no device that you can buy that's made out of molecular machines. But they're coming."

## MOLECULAR ARCHITECTS

In 1983, Sauvage's group was the first to create molecular interlocking chains and rings — called catenanes — which were the first steps to creating the connected parts needed for molecular motors. By creating interlocking rings, Stoddart noted at his press conference, Sauvage's group effectively invented a new way to bind molecules together — a mechanical bond, rather than a chemical one. "New bonds are few and far between. They are really the blue moons," Stoddart said.

Stoddart himself, in 1991, created the first molecular shuttle: a ring-shaped molecule threaded onto an 'axle', called a rotaxane. The

L: RSC; M: UNIV. GRONINGEN; R: VINCENT KESSLER/REUTERS



ring could shunt back and forth between two sites on the axle, which was capped at each end by stoppers, and Stoddart and other chemists worked out how to control that process, using changes in acidity, light or temperature.

Since then, Stoddart's team has used similar rotaxanes to make a molecular 'lift', which can raise itself (by less than a nanometre) above a surface, and an artificial 'muscle', in which rotaxanes bend a thin sheet. The researchers have also used millions of rotaxanes to make a high-density memory device — in which the shuttles flick from an 'on' state to an 'off' state.

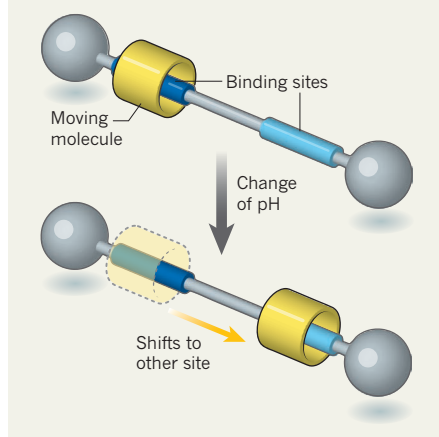
And in 1999, Feringa was the first to develop a synthetic molecular motor — a single molecule with paddle units connected by a carbon-carbon double bond. The paddles rotated, and kept on spinning, when the bond was broken with light. Feringa showed that the motors could have macroscale effects, such as rotating a glass rod sitting on top of them. Perhaps most famously, Feringa has also created a four-wheel-drive 'nanocar' out of the motors.

### WIDER IMPACTS

The Nobel prizewinners' work — and other chemists' nanomachines — have also had an impact on researchers' understanding of nature, Astumian says. In particular, the artificial systems have helped to demonstrate that all chemically powered molecular

### NANOMACHINES

Mechanisms the size of molecules are governed by the rules of chemistry, rather than Newtonian mechanics. An example is this switch called a rotaxane. A ring-shaped molecule is threaded onto a linear molecule and shifts between two binding sites as the acidity of the surrounding solution is altered.



machines, whether synthetic or biological, work on the same principle: they selectively harvest the random jiggles of Brownian motion, rather than push against them.

Asked by reporters at the Nobel press conference whether his machines would find a

use, Feringa likened the creators of minuscule machines to the Wright brothers, who made their maiden flight in a powered aircraft more than 100 years ago. "People were saying, why do we need a flying machine? Now we have a Boeing 747 and an Airbus. That's a little bit how I feel. The opportunities are great."

During his own press conference, Stoddart also took political swipes, both at recent UK anti-immigration rhetoric and at US Presidential candidate Donald Trump. He said that his old country, the United Kingdom, was "in a real mess because it thinks it can raise borders to people coming in". And referring to Trump's comment in his first debate with Hillary Clinton that not paying federal taxes would be "smart", Stoddart said that one-third of his Nobel earnings would go to taxes, because, he said, "I am not smart". ■

### CORRECTIONS

The News story 'Ukraine embraces solar and wind power' (*Nature* **537**, 598; 2016) gave the wrong year for the annexation of the Crimean peninsula. It happened in 2014. And the News story 'Medical award for cell recycling' (*Nature* **538**, 18–19; 2016) gave the wrong affiliation for Hitoshi Nakatogawa — he is at the Tokyo Institute of Technology.

# SEEING DEADLY MUTATIONS IN A NEW LIGHT

*How one of the largest genome resources in the world has quietly been changing scientists' understanding of human genetics.*

BY ERIKA CHECK HAYDEN

**L**urking in the genes of the average person are about 54 mutations that look as if they should sicken or even kill their bearer. But they don't. Sonia Vallabh hoped that D178N was one such mutation.

In 2010, Vallabh had watched her mother die from a mysterious illness called fatal familial insomnia, in which misfolded prion proteins cluster together and destroy the brain. The following year, Sonia was tested and found that she had a copy of the prion-protein gene, *PRNP*, with the same genetic glitch — D178N — that had probably caused her mother's illness. It was a veritable death sentence: the average age of onset is 50, and the disease progresses quickly. But it was not a sentence that Vallabh, then 26, was going to accept without a fight. So she and her husband, Eric Minikel, quit their

respective careers in law and transportation consulting to become graduate students in biology. They aimed to learn everything they could about fatal familial insomnia and what, if anything, might be done to stop it. One of the most important tasks was to determine whether or not the D178N mutation definitively caused the disease.

Few would have thought to ask such a question in years past, but medical genetics has been going through a bit of soul-searching. The fast pace of genomic research since the start of the twenty-first century has packed the literature with thousands of gene mutations associated with disease and disability. Many such associations are solid, but scores of mutations once suggested to be dangerous or even lethal are turning out to be innocuous. These sheep in wolves'

clothing are being unmasked thanks to one of the largest genetics studies ever conducted: the Exome Aggregation Consortium, or ExAC.

ExAC is a simple idea. It combines sequences for the protein-coding region of the genome — the exome — from more than 60,000 people into one database, allowing scientists to compare them and understand how variable they are. But the resource is having tremendous impacts in biomedical research. As well as helping scientists to toss out spurious disease-gene links, it is generating new discoveries. By looking more closely at the frequency of mutations in different populations, researchers can gain insight into what many genes do and how their protein products function.

ExAC has turned human genetics upside down, says geneticist David Goldstein of

ILLUSTRATION BY DARREN HOPES





Columbia University in New York City. Instead of starting with a disease or trait and working backwards to find its genetic underpinnings, researchers can start with mutations that look like they should have an interesting effect and investigate what might be happening in the people who harbour them. “This really is a new way of working,” he says.

ExAC is also providing better information for families facing genetic diagnoses. D178N, for example, was strongly suspected of causing prion disease because it had been seen in several people with the condition and seldom elsewhere. But before ExAC, no one really had the power to see just how rare it was. If it shows up in people more frequently than prion disease does, that would mean Vallabh's risk of getting the disease is much lower than predicted.

“We needed to find out if this mutation had ever been seen in a healthy population,” Minikel says.

#### DATA GATHERING

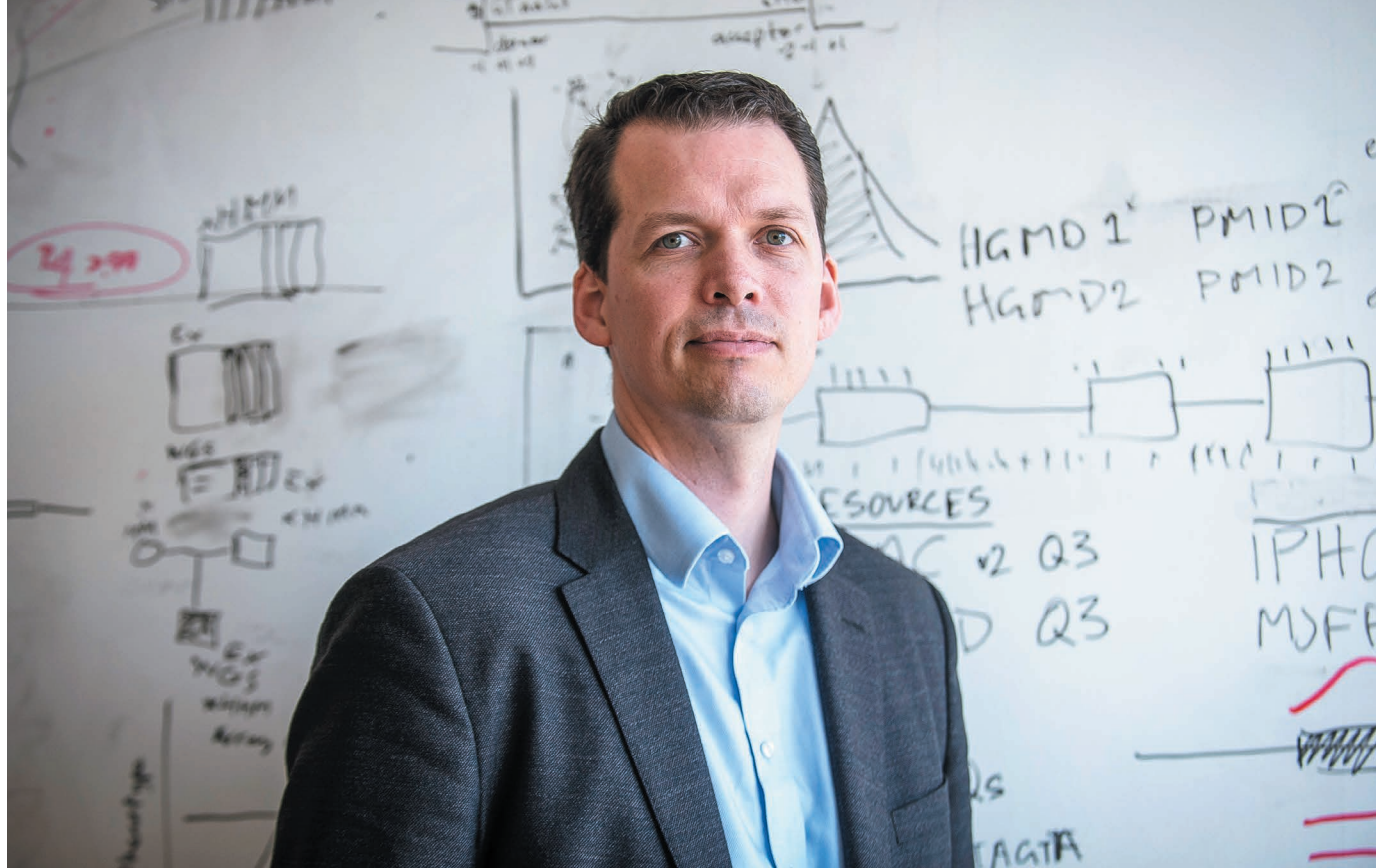
ExAC was born of frustration. In 2012, geneticist Daniel MacArthur was starting his first laboratory, at Massachusetts General Hospital (MGH) in Boston. He wanted to find genetic mutations that caused rare muscle diseases, and needed two things: genome sequences from people with these disorders, and genome sequences from people without them. If a mutation was more common in people with a disorder than in healthy controls, it stood to reason that the mutation was a likely cause.

The problem was that MacArthur couldn't find enough sequences from unaffected

people. He needed lots of exomes, and although researchers had been sequencing them by the thousands, existing data sets weren't large enough. No one had pulled enough together into one combined, standardized resource.

So MacArthur started asking his colleagues to share their data with him. He was well suited to the task: an early adopter of social media, his lively blog posts and acerbic Twitter feed had made him unusually popular and authoritative for a young scientist. He also had a position with the Broad Institute in Cambridge, Massachusetts, a genome-sequencing powerhouse. MacArthur convinced researchers to share data from tens of thousands of exomes with him; most were in some way connected to the Broad.

All that remained was to analyse the data, but that was no trivial task. Although the genes had



Daniel MacArthur convinced researchers to share genetic data on tens of thousands of people.

been sequenced, the raw data had been analysed using different types of software — including some that were out of date. If one individual in the collection showed a rare mutation, it could be real — or it could be an artefact of how different programs ‘called’ the bases within, judging whether they were As, Cs, Ts or Gs. MacArthur needed something that would standardize this gigantic data set. The Broad had developed genome-calling software, but it wasn’t up to the task of churning through the tremendous amount of data included in ExAC. So MacArthur’s team worked closely with the Broad programmers to test the software and scale up its abilities. “That was a pretty horrific 18 months,” MacArthur recalls. “We ran into every obstacle imaginable and had nothing to show for it.”

## PERSONAL STAKE

While this was going on, in April 2013, Vallabh was learning how to work with stem cells at MGH while Minikel studied bioinformatics. Minikel met MacArthur for lunch and explained his and Vallabh’s curiosity about whether D178N existed in healthy people. He admits to being a bit star-struck by MacArthur’s reputation. “I thought if I could get him to think about my problem for half an hour, that would probably be the most important thing that happened in my whole month,” Minikel says. The pair went upstairs to MacArthur’s lab, where bioinformatician Monkol Lek ran a search on the ExAC data that had been analysed so far — about 20,000 exomes. They didn’t see Vallabh’s mutation. That wasn’t good news, but, optimistic about exploring the data further, Minikel joined MacArthur’s lab.

By June 2014, MacArthur’s team and its

collaborators had a data set that they were confident in — exomes from 60,706 individuals representing various ethnic groups, who met certain thresholds for health and consent. They released ExAC that October at the annual meeting of the American Society of Human Genetics (ASHG), in San Diego, California. Immediately, researchers and physicians recognized that the data could help to recast their understanding of genetic risks.

Many disease-association studies, particularly in recent years, have identified mutations

**“IF YOU HAVE A GENETIC RISK THAT YOU BELIEVE IS PREDICTING DISEASE BUT ISN’T, YOU CAN END UP DOING DRASTIC THINGS.”**

as pathogenic simply because scientists performing analyses on a group of people with a disorder found mutations that looked like the culprit, but didn’t see them in healthy people. But it’s possible that they weren’t looking hard enough, or in the right populations. Baseline ‘healthy’ genetic data has tended to come mainly from people of European descent, which can skew results.

In August this year, MacArthur’s group published<sup>1</sup> its analysis of ExAC data in *Nature*, revealing that many mutations thought to be harmful are probably not. In one analysis, the group identified 192 variants that had previously been thought to be pathogenic,

but turned out to be relatively common. The scientists reviewed papers about these variants, looking for plausible evidence that they actually caused disease, but could find solid evidence for only nine of them. Most are actually benign, according to standards set by the American College of Medical Genetics and Genomics, and many have now been reclassified as such.

Similar work promises to have direct impacts on medical practice. In a companion paper<sup>2</sup>, geneticist Hugh Watkins of the University of Oxford, UK, looked at genes associated with certain types of cardiomyopathy that cause gradual weakening of the heart muscle. Undetected, they can lead to sudden death, and it has become fairly common to check relatives of people with the conditions for genetic mutations associated with them. Those found to have a genetic risk are sometimes counselled to get an implanted defibrillator, which delivers electrical shocks to the heart if it seems to be beating abnormally. Watkins checked the ExAC database for information on genes that have been associated with these heart conditions, and found that many mutations are much too common among healthy people to be pathogenic. About 60 genes had been implicated as harbouring pathogenic mutations that cause one form of the disease; Watkins’ analysis revealed that 40 of these probably bear no link.

This was troubling. “If you have a genetic risk that you believe is predicting disease but isn’t, you can end up doing drastic things that can harm someone,” says Watkins.

Even some of the mutations that seem to be reliably linked to disease aren’t a sure bet — such as those in *PRNP*. There are definitely mutations in the gene that cause the disease, but some



variants might not be pathogenic or might elevate the risk only slightly (see ‘The deadly mutations that weren’t’). To find out the status of D178N, Vallabh and Minikel gathered genetic data from more than 16,000 people who had been diagnosed with prion diseases, and compared them with data from almost 600,000 others, including the ExAC participants<sup>3</sup>.

The pair found that 52 people in ExAC had *PRNP* mutations that have been linked to prion diseases, but based on the prevalence of the disease, they would have expected to see maybe two. Minikel calculated that some of these supposedly lethal mutations elevated a person’s risk of prion disease slightly; some seemed not to be linked to prion disease at all.

This work provided insight for people such as Alice Uflacker. In 2011, Uflacker’s father, Renan, died from Creutzfeldt–Jakob disease, a prion illness that causes rapid mental and physical deterioration. He was 62. Alice found out that she carried a mutation in *PRNP* called V210I, which had been linked to her father’s disease in previous studies. Three years later, she learned from Minikel that the mutation confers, at most, a small risk of disease. The information was helpful, and the result made sense; her grandmother had lived to 93 despite having the same mutation.

Vallabh and Minikel would find no such relief, however. D178N was absent from the other genomes they looked at, and is still highly likely to cause prion disease. Minikel and Vallabh had already begun to suspect as much, as Minikel dug into the data. “All along the way was gradual confirmation of what we were assuming anyway,” Minikel says. “There wasn’t any moment where we said, ‘Ah, this is the worst news.’ We’d already gotten the worst news.”

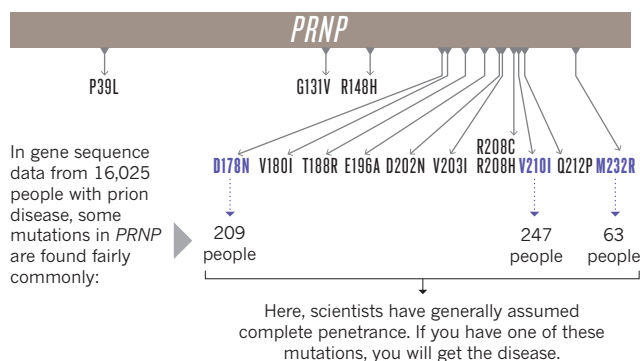
## HUMAN KNOCKOUTS

ExAC is revealing a lot about genes through the frequency of mutations. MacArthur and his team found<sup>1</sup> 3,200 genes that are almost never severely mutated in any of the ExAC genomes — a signal that these genes are important. And yet 72% of them have never before been linked to disease. Researchers are eager to study whether some of these genes play unappreciated parts in illness.

Conversely, the group has found nearly 180,000 instances of mutations so severe that they should render their protein products completely inactive. Scientists have long studied genes by knocking them out in animals such as mice, so that they don’t work. By

## THE DEADLY MUTATIONS THAT WEREN’T

Prion diseases are rare neurodegenerative disorders caused by misfolded prion proteins. About 63 mutations in the gene *PRNP* have been linked to them. But until now it has been difficult to estimate how likely it is that a given variant will result in disease, a measure known as penetrance. Data compiled by the Exome Aggregation Consortium (ExAC) can help.



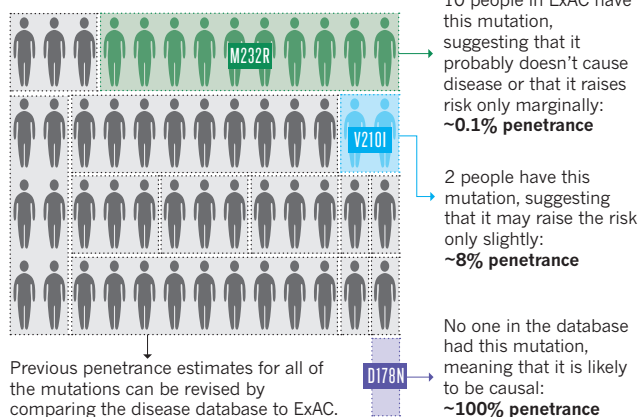
## ExAC DATABASE STUDY

Total prion disease occurrence: 2 people in every 1,000,000 per year.

ExAC contains the protein-coding sequences of 60,706 people.

Number of people with *PRNP* mutations expected in ExAC: 1.7

Actual number of people with mutations: 52



looking at the symptoms that develop, they can study what the genes do. But that has never been possible in humans. Now, researchers are eager to study these natural human knockouts to understand what they can reveal about how diseases develop or may be cured. MacArthur and other researchers are gearing up to prioritize which human knockout genes to study and how best to contact the people carrying them for further study.

But it will have to wait until he completes the second phase of ExAC. Due to be unveiled at the ASHG meeting in Vancouver, Canada, this month, it will double the data set's size to 135,000 exomes and include some 15,000 whole-genome sequences, which should allow researchers to explore mutations in regulatory regions of the genome that are not captured by exome sequencing.

ExAC is quietly becoming a standard tool in

medical genetics. Clinical labs around the world now check it before telling a patient that a particular glitch in their genome might be making them ill. If the mutation is common in ExAC, it's unlikely to be harmful. Geneticist Leslie Biesecker at the US National Human Genome Research Institute in Bethesda, Maryland, says that his lab uses ExAC daily in patient care. “It’s a critical factor that we take into consideration for every variant,” he says. He and other geneticists are now embarking on a painstaking reckoning with the genetics literature that will probably take years.

ExAC has also driven home a point that Goldstein and other researchers have made repeatedly: that failing to include people from Asian, African, Latino and other non-European ancestries is holding back understanding of how genes influence disease by limiting the view of human genetic diversity (see page 161). There is now a fresh impetus to include under-represented groups in planned studies linking genetics and health information on large numbers of people, such as the US Precision Medicine Initiative.

For Vallabh and Minikel, ExAC provided a disheartening confirmation, but also some promising insight. Minikel’s studies have identified<sup>3</sup> three people in ExAC with mutations that should silence one of the two copies of the prion protein gene. If they can live with a limited amount of functioning protein, perhaps a drug could be made that would silence the defective protein in Vallabh, preventing prion aggregation and disease progression without dangerous side effects. Minikel got in touch with one of the individuals, a man in Sweden,

who agreed to donate some cells for research. Minikel and Vallabh have now joined the lab of biochemist Stuart Schreiber at the Broad Institute, where they are working full-time to find candidate drugs to treat prion disease.

The couple exemplifies the challenge of translating ExAC data into real medical benefits. “We can’t go back from this,” Vallabh says. “We have to go through it.” Their situation couldn’t be more illustrative of what is at stake: Vallabh is now 32 — just 20 years younger than her mother was when she died. She has no time to waste. ■ SEE EDITORIAL P.140

**Erika Check Hayden** writes for Nature from San Francisco.

1. Lek, M. et al. *Nature* **536**, 285–291 (2016).
2. Walsh, R. et al. *Genet. Med.* <http://dx.doi.org/10.1038/gim.2016.90> (2016).
3. Minikel, E. V. et al. *Sci. Transl. Med.* **8**, 322ra9 (2016).



# The troubled minds of migrants

THE REFUGEES AND MIGRANTS SURGING INTO EUROPE ARE SUFFERING VERY HIGH LEVELS OF PSYCHIATRIC DISORDERS. RESEARCHERS ARE STRUGGLING TO HELP.

*By Alison Abbott*

**O**n an ice-cold day in January, clinical psychologist Emily Holmes picked up a stack of empty diaries and went down to Stockholm's central train station in search of refugees. She didn't have to look hard. Crowds of lost-looking young people were milling around the concourse, in clothes too flimsy for the freezing air. "It struck me hard to see how thin some of the young men were," she says.

Holmes, who works at Stockholm's Karolinska Institute, was seeking help with her research — a pilot project on post-traumatic stress disorder (PTSD), which is all too common in refugees. She wanted to see whether they would be willing to spend a week noting down any flashbacks — fragmented memories of a trauma that rush unbidden into the mind and torment those with PTSD. She easily

found volunteers. And when they returned the diaries, Holmes was shocked to see that they reported an average of two a day — many more than the PTSD sufferers she routinely dealt with. "My heart went out to them," she says. "They managed to travel thousands of kilometres to find their way to safety with this level of symptoms."

Europe is experiencing the largest movement of people since the Second World War. Last year, more than 1.2 million people applied for asylum in the European Union — and those numbers underestimate the scale of the problem. Germany, which has taken in the lion's share of people, reckons that it received more than a million refugees in 2015, tens of thousands of whom have yet to officially apply for asylum. Most came from Syria, Afghanistan and Iraq. Many have experienced war, shock, upheaval and terrible journeys, and they often have poor physical health. The crisis has attracted global attention and sparked political tension as countries struggle to accommodate and integrate the influx.

What hasn't been widely discussed is the enormous burden of mental-health disorders in migrants and refugees. Clinical psychologist Thomas Elbert from the University of Konstanz in Germany is

JEROME SESSINI/MAGNUM



**Arrival in a foreign, hostile country causes many refugees great stress.**

conducting a local survey of refugees that suggests “more than half of those who arrived in Germany in the last few years show signs of mental disorder, and a quarter of them have a PTSD, anxiety or depression that won’t get better without help”. Previous research shows that refugees and migrants are also at a slightly increased risk of developing schizophrenia.

“It is a public-health tragedy — and it’s a scandal that it is not recognized as such, as a physical epidemic would be,” says epidemiologist James Kirkbride of University College London.

Doctors and researchers are starting to take action. Holmes and other psychologists and psychiatrists are working with refugees to develop practical, cheap and effective therapies for trauma-related disorders — therapies that could be quickly deployed on this group. Other scientists want to work with local refugees to understand more about how the different types of stresses they suffer play out in their brains, and to learn more about the basic biology of psychiatric disorders.

Scientists hope that their studies will help them to deal with other displaced populations, and help policymakers to accommodate the current influx. Politicians have been too slow to consider mental health when they call for refugees to integrate quickly, Elbert says. “It is illusory to think that people can learn a new language and find work when they can’t function properly mentally. If we want quick integration, we need an immediate plan for mental health.”

## MAKING A NEW LIFE

Amira is a clinical psychologist and a refugee from Syria. When the war there started, she worked in camps for Syrian refugees in Jordan. She saw people who had been physically attacked, women who had been raped and children who had been neglected. The symptoms of PTSD were clear, and she knows that many refugees have depression and anxiety, too. She asked that her real name not be used.

She arrived in Sweden at the end of December 2015, and wanted to help other refugees but was not allowed to work at first. She tried to make contacts in Stockholm and joined a language course for refugees; she felt very alone but carried on. Now, she has a 6-month position. “I met many children who have experienced war,” she says. “We feel sad, [about] how our children think and how they feel. I have a child and I try to protect him.”

Researchers already have a wealth of evidence about the mental health of migrant and refugee populations around the world. (The United Nations defines refugees as people fleeing armed conflict or persecution and migrants as people who choose to move to improve their lives. Asylum seekers are those seeking official refugee status; but sometimes different definitions are used.) A 2005 meta-analysis of studies performed mostly in northern Europe showed that first- and second-generation migrants were at much greater risk of schizophrenia than non-migrants — and that those from developing countries were more at risk than those from developed ones<sup>1</sup>.

A large cohort study published in March looked at 1.3 million people who had arrived in Sweden before 2011 (see ‘Migrant crisis’). Refugees had a threefold higher incidence of schizophrenia and other psychotic disorders than native-born Swedes, and a 66% higher incidence than migrants who were not refugees<sup>2</sup>. (The overall risk for refugees and migrants still remains comparatively low, at perhaps 2–3%.) Kirkbride, an author on the study, says that his team’s more recent analysis of UK migration data suggests that the level of increased risk of psychotic disorders may depend on how old people were when they migrated — with children potentially at greater risk.

Those who stand out most seem to be particularly vulnerable. The 2005 meta-analysis showed that black migrants in a mostly white population had an almost fivefold increased risk of psychotic disorders<sup>1</sup>. And the risk is higher for migrants living in neighbourhoods with a low proportion of residents from their own ethnic group compared with those surrounded by many of their own ethnicity<sup>3</sup>.

Psychiatrist Andreas Meyer-Lindenberg from the Central Institute

for Mental Health in Mannheim, Germany, is one of those trying to understand the brain mechanisms involved. He has already studied other populations with an above-average risk of psychosis, such as city dwellers<sup>4</sup> and ethnic minorities<sup>5</sup>. The work suggested that the brains of these people are overly sensitive to social stress, such as a stream of disapproving feedback.

Thanks to a grant awarded last month from the state government of Baden-Württemberg, Meyer-Lindenberg plans to extend his studies by recruiting 200 refugees and 200 people from the local community. The refugees will use smartphones to note their state of mind — such as feelings of suspiciousness — and they will later receive brain scans. The eventual aim is to find patterns in the data that indicate people with abnormal processing of social stress, who may therefore be at increased risk of mental illness. Jean-Paul Selten, a psychiatrist at Maastricht University in the Netherlands, is also exploring the poisonous nature of social stress. He proposes that pressures such as social exclusion raise the risk of psychosis by changing the brain’s sensitivity to the neurotransmitter dopamine<sup>6</sup>.

Germany’s integration plans, consolidated in a law that came into force in August, involve distributing refugees across the country to avoid the creation of large, isolated ethnic communities. That could be problematic if it increases people’s isolation, but Meyer-Lindenberg says it’s “actually a good policy” — because other people in the community get to know refugees, and this usually reduces xenophobia, another major source of social stress.

## “IT IS A PUBLIC-HEALTH TRAGEDY — AND IT’S A SCANDAL THAT IT IS NOT RECOGNIZED AS SUCH.”

Politicians consider integration to be essential for security, among other things. A handful of terrorist attacks in Europe over the past two years have been carried out by refugees or others with a migrant background who had been known to have a history of psychiatric problems. But doctors and researchers are extremely wary about making a link between refugees or migrants and terrorist acts, pointing out that very few of those with mental-health problems become violent, regardless of their origins. The security concern simply accentuates the need to help all those with mental-health problems across the population, they say.

## THE STRESS OF UPHEAVAL

Psychologists recognize three windows of extreme stress for refugees: the often violent traumas in their home countries that led to their flight; the journey itself; and the arrival, when people are thrust into a foreign country. “The latter ‘post-migration’ phase is becoming increasingly important,” says psychiatrist Malek Bajbouj from the Charité university hospital in Berlin. “Suddenly they realize they have lost everything, have no control over aspects of their lives and no social standing.”

In February, Bajbouj — who is of Syrian descent and speaks Arabic — and two colleagues from other departments opened a clearing centre for refugees with mental-health problems, the first of its kind in Germany. It’s a quiet building, a former hospital in the centre of Berlin — but already 1,500 troubled people have passed through its doors. “Refugees may arrive in Germany with great hope, but then find themselves stuck for months in camps with no apparent prospects,” he says. “When we ask them what their greatest stressors are, they typically refer not to their traumatic memories, but to their current frustrations.”

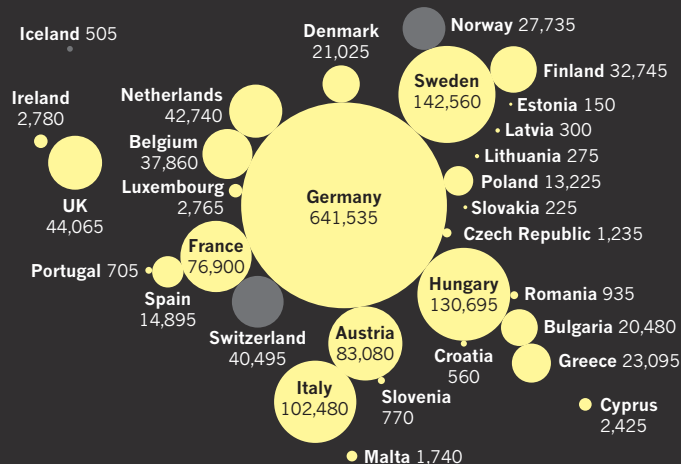
The biggest challenge for Bajbouj and others is the sheer volume of people in need of help. They must be assisted quickly and cheaply, in ways that take the pressure off overstretched health professionals. At the clearing centre, three psychiatrists assess visitors rapidly, categorizing them into those who require low-level or more-intensive psychiatric help and those who can be aided by social workers. A lot of effort goes into teaching about stress management and the science behind mental health. “Some people from rural areas hold the Djinn responsible

# Migrant crisis

## NUMBERS

More than 1.4 million people applied for asylum in Europe for the first time in the second half of 2015 and first half of 2016, with the biggest share in Germany.

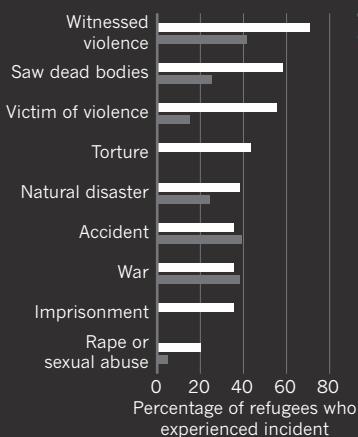
● EU ● Non-EU



## TRAUMA

Many refugees in Germany have experienced traumatic incidents; around half suffer from mental illnesses such as post-traumatic stress disorder or depression.

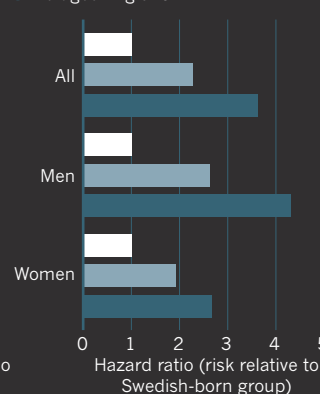
● Adults ● Children



## PSYCHOSES

A study of 1.3 million people in Sweden<sup>2</sup> showed that migrants and refugees have a higher incidence of schizophrenia and other non-affective psychotic disorders.

● Swedish born  
● Non-refugee migrant  
● Refugee migrant



for their moods,” says Bajbouj, “and we teach them that symptoms like sleeplessness and depression are biologically based and can be treated.”

Elbert wants to see similar triage systems put in place across Germany. In a paper to be published next month, he and a group of colleagues call for a three-tiered approach. Refugees would initially be helped by bilingual laypeople — ideally migrants or refugees themselves — who are trained to guide people through the German health system (tier one) or to offer trauma counselling (tier two). Those in most need would progress to tier three: qualified psychologists or psychiatrists.

## PEACE OF MIND

Training laypeople seems to work in emergency situations. Elbert, together with Sarah Ayoughi, a clinical psychologist from the psychosocial-care organization Ipsos, carried out a randomized controlled study of people with mental-health conditions in north Afghanistan who received psychosocial counselling — a type of talking therapy — conducted by local physicians who had no previous education in psychology or psychiatry, but who were specially trained for the trial. Just 5–8 sessions improved symptoms of depression and anxiety for up to 3 months<sup>7</sup>.

And several studies, including a 2011 randomized controlled trial of former child soldiers in northern Uganda<sup>8</sup>, show that an approach called narration exposure therapy (NET), carried out by trained lay counsellors, can reduce the severity of PTSD symptoms. Elbert started developing NET with his wife Maggie Schauer, also a clinical psychologist at the University of Konstanz, when they were working with refugees in Kosovo in the late 1990s. It exploits new understanding of how memories are linked with fear circuitry in the brain. A traumatized person works with a therapist or counsellor to construct a narrative of their lives and anchor their traumatic experiences in the correct time and place.

Pragmatic as the three-tiered approach may sound, it won't be simple to introduce in Germany. Professional associations are resistant to allowing people without formal qualifications to help out with psychotherapies, and various regulations could get in the way.

But while the federal government ponders what to do, some programmes are starting up with regional government support. Schauer has received €100,000 (US\$112,000) to test whether NET works as well on refugees in Germany as it has in war-torn countries. And Ayoughi is organizing the training of refugees in Erfurt in Germany, with additional support from the Google Foundation.

Bajbouj thinks that the political desire to get refugees into the workforce fast may end up easing the way for more relaxed rules about psychotherapy. And there is another way to deliver inexpensive mental-health care: through the Internet and apps. He is developing an

Arabic-language version of the smartphone app PTSD Coach, which provides education, a personalized emergency plan, self-assessment and 25 different techniques to regulate stress. He is testing it in the Arab Outpatient Centre he opened at the Charité in 2008.

In Stockholm, Holmes also hopes that technology can help. The aim of her work is to test whether it's possible to subdue PTSD-linked emotional flashbacks if a person immediately plays a video game on their phone that competes for cognitive space in the brain — a technique that she has seen work in laboratory tests<sup>9</sup>. “The important thing now is to develop simple new approaches to therapy that can be scaled up, and to prove that they help,” she says.

Sweden, which has taken in a relatively large number of refugees, is also starting mental-health programmes. Early this year, local authorities rolled out a plan to make it easier for refugees to access support: health checks will include more questions about states of mind, and those recognized as being in need will be channelled towards psychological or psychiatric support.

The flow of refugees and migrants has eased this year, in part because Turkey has agreed to take back those who illegally entered EU countries from there. But people keep coming. This August, more than 18,000 refugees entered Germany to seek asylum. And even if the current crisis eases, conflict, poverty, natural catastrophe and climate change will inevitably drive fresh waves of migration around the world. “We've learnt lessons about mental health from crises in war-torn countries,” says Ayoughi, “and we can apply these in the refugee crisis in Europe now if we get the support.” Then, perhaps, lessons learnt in Europe could feed back to war zones.

Bajbouj has been calling for a ‘migration think-tank’, a permanent institution in Germany where scientists of different disciplines can come together to work out what needs to be done. “The challenges are not just about mental health, but about education, integration into the work force and much more,” he says. “But mental health impacts everything.” ■

**Alison Abbott** is Nature's senior European correspondent.

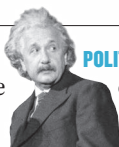
1. Cantor-Graae, E. & Selten, J.-P. *Am J Psychiatry* **162**, 12–24 (2005).
2. Hollander, A.-C. *et al. Br. Med. J.* **352**, i1030 (2016).
3. Kirkbride, J. B., Jones, P. B., Ullrich, S. & Coid, J. W. *Schizophr. Bull.* **40**, 169–180 (2014).
4. Lederbogen, F. *et al. Nature* **474**, 498–501 (2011).
5. Akdeniz, C. *et al. JAMA Psychiatry* **71**, 672–680 (2014).
6. Selten, J.-P., van der Ven, E., Rutten, B. P. F. & Cantor-Graae, E. *Schizophr. Bull.* **39**, 1180–1186 (2013).
7. Ayoughi, S., Missmahl, I., Weierstall, R. & Elbert, T. *BMC Psychiatry* **12**, 14 (2012).
8. Ertl, V., Pfeiffer, A., Schauer, E., Elbert, T. & Neuner, F. *JAMA* **306**, 503–512 (2011).
9. James, E. L. *et al. Psychol. Sci.* **26**, 1201–1215 (2015).



# COMMENT

**CITIES** To inform policy, urban scholarship must get organized and funded **p.165**

**HISTORY** A biography of Enrico Fermi, Italy's fallible atomic physicist **p.168**



**POLITICS** The causes Einstein championed offer a window on his time **p.170**

**OBITUARY** Roger Yonchien Tsien, fluorescent-biology pioneer, remembered **p.172**

CYRUS MCCORMICK/DENVER POST/GETTY



Certain drugs may be less effective, or even unsafe, in some populations because of genetic differences.

## Genomics is failing on diversity

An analysis by **Alice B. Popejoy** and **Stephanie M. Fullerton** indicates that some populations are still being left behind on the road to precision medicine.

A 2009 analysis revealed that 96% of participants in genome-wide association studies (GWAS) were of European descent<sup>1</sup>. Such studies scan the genomes of thousands of people to find variants associated with disease traits. The finding prompted warnings that a much broader range of populations should be investigated<sup>2</sup> to avoid genomic medicine being of benefit merely to “a privileged few”.

Seven years on, we’ve updated that

analysis. Our findings indicate that the proportion of individuals included in GWAS who are not of European descent has increased to nearly 20%. Much of this rise, however, is a result of more studies being done in Asia on populations of Asian ancestry. The degree to which people of African and Latin American ancestry, Hispanic people and indigenous peoples are represented in GWAS has barely shifted.

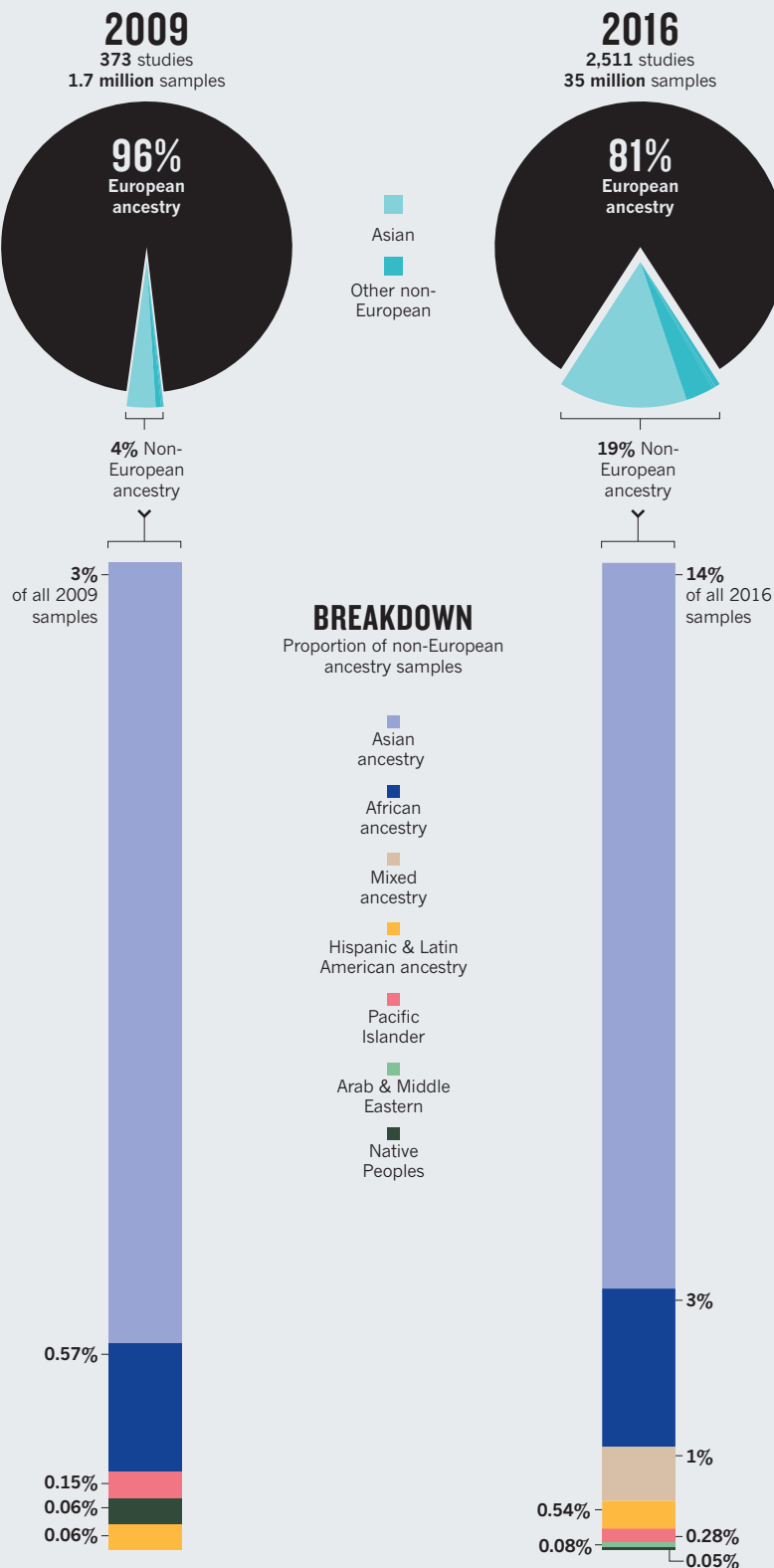
Thus, more than 20 years after the

US National Institutes of Health (NIH) mandated the inclusion of diverse participants in the biomedical research it funds, GWAS funded by the NIH and other sources are continuing to miss a vast portion of the world’s genetic variation.

Over the past decade, GWAS have been the preferred tool for discovering the genetic factors involved in common diseases. Tens of thousands of significant associations between genetic variants and biological traits have ►

## PERSISTENT BIAS

Over the past seven years, the proportion of participants in genome-wide association studies (GWAS) that are of Asian ancestry has increased. Groups of other ancestries continue to be very poorly represented.



Terms for ethnicity are those used in the GWAS Catalog. Some have changed between 2009 and 2016 as sampling has increased. Samples of European origin have the most specific descriptions of population ancestry.

► now been found, and many of these associations have helped geneticists to uncover biological mechanisms underpinning conditions from diabetes to schizophrenia.

The most comprehensive, publicly accessible summary of human genetic association research is the GWAS Catalog ([www.ebi.ac.uk/gwas](http://www.ebi.ac.uk/gwas)) produced by the US National Human Genome Research Institute in partnership with the European Bioinformatics Institute. Every week, the curators of the catalogue receive automatic alerts of any new English-language GWAS reported in PubMed. These studies are then put through two rounds of data extraction and validation before being added to the catalogue. Among the data extracted from each study are the race, ethnicity or ancestry (as described by the authors of the study) of the subjects whose samples were analysed.

### DATA GATHERING

To determine ancestry, we analysed the sample descriptions included in the GWAS Catalog with an approach similar to that used in 2009 (see Supplementary Information; [go.nature.com/2dv2faf](http://go.nature.com/2dv2faf)).

As of August, 2,511 studies involving nearly 35 million samples were included in the GWAS Catalog. This is a more than 2,000% increase in sample number from the 2009 analysis (which looked at roughly 1.7 million samples across 373 independent studies<sup>1</sup>).

We found considerable heterogeneity in descriptions. For example, 26 terms, including 'black cases' and 'sub-Saharan African', were used to describe people of African ancestry. The most geographically specific and informative descriptions were those used for samples of European origin, as previous studies have shown<sup>3</sup>.

During the past seven years, the proportion of samples used in catalogued GWAS from participants who are not of European descent has increased fivefold (see 'Persistent bias'). Yet nearly 78% of this growth is due to an increase in the number of samples from Japan, China, Korea, India and other populations from east Asia, south Asia and southeast Asia.

Together, individuals of African and Latin American ancestry, Hispanic people (individuals descended from Spanish-speaking cultures in central or South America living in the United States) and native or indigenous peoples represent less than 4% of all samples analysed. Collectively, these are the most vulnerable and traditionally underserved populations in many of the world's richest nations.

The proportion of samples from individuals of African ancestry has increased by 2.5%, and the proportion of people of Hispanic or of Latin American ancestry by around 0.5%. In the case of indigenous peoples (including Native Americans, Australian Aboriginals



and Pacific Islanders), representation has decreased slightly since 2009.

By looking up GWAS involving only Asian participants in PubMed (349 studies), we found the institution of the first author of each study. Around 93% of these studies were conducted in Asian countries. That the number of GWAS involving local populations has risen so much in Asia is heartening. But with such a large increase overall in the number of GWAS performed in the past seven years, the lack of growth in representation from other populations is remarkable and deeply disconcerting.

Of course, our analysis does not account for the resampling of data sets across independent studies. Information from some cohorts in publicly available databases has been used multiple times for different GWAS (see Supplementary Information). So the numerous samples of European ancestry used in GWAS could come from a smaller number of actual individuals. Yet if European-ancestry data sets are resampled more often than others, this in itself reflects population-specific differences in research effort.

### WHY THE BIAS?

The continuing European bias in GWAS is likely to be the result of logistical, systemic and historical factors.

The more populations that are included in a study, the more variables there are to control for. In trying to keep things as simple as possible, geneticists probably favour the use of existing cohorts, such as that of the Framingham Heart Study, or other large data sets generated by well-established medical centres.

Such organizations collect samples and information from people in the same geographic location, who are presumed to be exposed to shared environmental factors, using uniform collection practices. But for various reasons, some populations are easily bypassed. People may have limited access to certain medical centres, for example, or, for cultural or historical reasons, elect not to contribute their samples to research.

Genotype and phenotype information from diverse populations is available. Researchers using NIH funding are required to submit any such information they have collected to dbGaP, a public database of genotypes and phenotypes. Analogous recommendations are made by other major biomedical funders outside the United States. In Europe, geneticists are encouraged to share similar data through the European Genome-phenome Archive (EGA). Yet for various reasons (such as the difficulties of getting certain kinds of studies funded, a preference for larger sample sizes, a perception that the analysis will be simplified by using data from one ancestry group or a lack of awareness of the diversity of data sets available) geneticists seem to be preferentially



A study of Greenlandic Inuits revealed a previously missed genetic variant associated with height.

using cohorts of European ancestry.

Repeated sampling is almost certainly exacerbating the problem. Indeed, to some degree, the over-representation of people of European ancestry in GWAS may be a legacy of earlier biases.

### WHAT'S MISSED

Irrespective of what's driving it, the continued under-representation of populations of mixed ancestry or of people whose ancestry is not European is a problem.

Until they are able to conduct amply powered GWAS on each major ancestral population across the world, geneticists will continue to miss important information about disease biology. They won't know how many of the thousands of associations between variants and diseases, and between variants and responses to drugs, observed in populations of European ancestry replicate in other groups. And opportunities will be missed to discover new associations with disease traits in other populations.

For example, for 25% of the variants in European Americans that GWAS have identified as being associated with body mass index, type 2 diabetes and lipid levels, the strength of the association differs in at least one out of five populations of non-European ancestry<sup>4</sup>. This means that a variant that is associated with

diabetes may confer a different risk of disease in someone of European ancestry than in, say, an individual of African ancestry.

Likewise, population-specific differences in the frequencies of variants associated with drug metabolism may mean that certain drugs will be safer and more effective in some populations than in others. The *CYP2D6* gene, for instance, is involved in the metabolism of many commonly prescribed drugs, including tamoxifen, which is used to treat breast cancer. More than 100 different variants of this gene (alleles) — many of which affect an individual's ability to safely digest and use a drug<sup>5</sup> — occur at different frequencies in different populations.

Several associations between drug responses and clinically relevant genetic variants have already been identified with GWAS. In some cases in which the effect sizes are large, significant results have been found with as few as 51 cases and 282 controls<sup>6</sup>. (In this case, patients had different reactions to the lipid-lowering drug simvastatin.) Although physicians must weigh the costs and benefits of using pharmacogenetic testing to guide prescription and dosage decisions for individual patients, these findings suggest that the small samples that have already been collected from under-represented populations could yield leads that have not been identified in populations of European ancestry.

Conducting analyses in other populations is also crucial for assessing the accuracy and broader relevance of a finding. It is possible,

*“European ancestry in GWAS may be a legacy of earlier biases.”*

for example, that associations between certain disease traits and variants found in European populations that cannot be replicated in other populations are actually false positives. In fact, the analysis of a broader representation of populations can reveal insights that would have otherwise been missed.

A genome-wide scan in a Greenlandic Inuit population, for example, found last year that a single-nucleotide polymorphism (SNP) in a fatty-acid enzyme affects height in both this population and Europeans<sup>7</sup>. The authors suggest that previous GWAS may have missed this variant because of its low frequency in Europeans (0.017 compared to 0.98 in the Greenlandic Inuit population) — even though it has a much greater effect on height than others previously identified through GWAS.

### NEW DIRECTIONS

Increasingly, the sequencing of whole genomes and whole exomes (that is, the complete set of protein-coding genes) are beginning to be used more widely for discovery as costs fall. These may prove more fruitful than GWAS for individual-level diagnosis and treatment. Certainly, they are better suited to revealing rare variations that are clinically informative. (GWAS identify known genetic markers associated with a trait, but not necessarily the mutations that cause the disease.)

Studies that use these new approaches have been slightly more successful than GWAS at recruiting a greater diversity of populations. For example, the international Exome Aggregation Consortium hosts data on genetic variants from more than 60,000 samples, of which 8.6% are from people of African ancestry, 9.5% are from people of Latin American ancestry, and 60.4% are from people of European ancestry<sup>8</sup> (see page 154). The remaining samples (21.5%) are from south Asia, east Asia and the Middle East. Similarly, the Trans-Omics for Precision Medicine whole-genome sequencing project of the US National Heart, Lung and Blood Institute is growing and currently holds 62,000 samples, of which 50% are from European Americans, 30% are from African Americans, 10% are from Hispanics or Latin Americans, and 8% are from Asians.

Often, large sample sizes are needed to uncover rare genetic variants associated with disease traits. In fact, this realization — from the first generation of exome discovery studies — is driving new interest in ultracheap genotyping arrays (collections of targeted fragments of DNA). Using such arrays, geneticists can speed up the sequencing process and analyse many targeted samples in one go. Exome sequencing combined with the use of genotyping arrays is likely to be the favoured approach over the next decade. Nonetheless, GWAS remains a useful precursor to such studies, as well as to those involving whole-genome sequencing.

And emerging data indicate that inequalities in health care are being exacerbated by findings from whole-exome and -genome sequencing, despite their greater sample diversity compared with GWAS. Patients of African and Asian ancestry are currently more likely than those of European ancestry to receive ambiguous genetic test results after

**“Historical, cultural, scientific and logistical factors are sustaining an embarrassing bias in genomics.”**

exome sequencing, or be told that they have variants of unknown significance<sup>9</sup>. Furthermore, patients of African ancestry are more likely than those of European ancestry

to be wrongly told that a mutation they carry increases their risk of developing a life-threatening heart condition known as hypertrophic cardiomyopathy<sup>10</sup>. Had more ethnically diverse controls been included in the candidate-gene studies that identified these associations, population-specific differences in the frequency of presumed disease-causing variants would have revealed a false positive at the outset.

### WHAT NOW?

The message being broadcast by the scientific and medical genomics community to the rest of the world is currently a harmful and misleading one: the genomes of European descendants matter the most.

Certain efforts, combined with newer data-gathering initiatives, can help to move the needle in the right direction. Some investigators in genomics focus exclusively on diverse populations. For instance, landmark trans-ethnic studies have identified genes associated with traits such as diabetes, levels of lipids and other metabolites, prostate cancer and gene expression<sup>11</sup>. Also, various ventures aim to boost genomics studies in under-represented populations worldwide. The Human Heredity and Health in Africa Consortium, for example, was established by the NIH and the Wellcome Trust in London in 2012 to help build infrastructure and genomics expertise across Africa.

In our view, more fundamental changes are needed — both top-down and bottom-up. Funding agencies should develop financial incentives for the creation of diverse cohorts of study participants. One way for them to do this would be to prioritize grant requests that propose investigations in populations of non-European (and especially of African) ancestry. Given limited budgets, this may need to happen hand in hand with a reduction in the funding of research on existing cohorts of European ancestry for traits and diseases that have been relatively well characterized. (Around 850 genetic associations with height have now been reported by roughly

30 independent GWAS — the vast majority of which have been conducted using individuals of European ancestry.)

Further, all genomics researchers need to recognize the importance of studying under-represented populations to ensure that the benefits of research are distributed fairly and to maximize the potential for discovery. On a practical level, training programmes and new infrastructure, such as good health-care clinics that provide genetic testing in predominantly black or Hispanic neighbourhoods, could enhance trust and allow people to engage in projects as stakeholders rather than as study participants.

A culture shift is required at every level. Efforts to recruit participants for biomedical research in under-represented communities have been most successful when conducted by investigators of concordant racial or ethnic background, and in partnership with institutions trusted by those communities<sup>12</sup> — such as historically black colleges and universities in the United States.

Indeed, to a large extent, the persistent bias in sampling in genomics mirrors the employment trends evident in biomedical institutions worldwide. In the United States in 2012, less than 4% of the tenured and tenure-track faculty members in research-intensive biomedical departments were African American, Hispanic or Native American<sup>13</sup>.

A complex web of historical, cultural, scientific and logistical factors is sustaining an embarrassing bias in genomics. Before precision medicine takes hold in clinical practice, we must correct its course. ■

**Alice B. Popejoy** is a PhD candidate at the Institute for Public Health Genetics (IPHG) at the University of Washington, Seattle, USA. **Stephanie M. Fullerton** is associate professor of bioethics and humanities at the University of Washington, Seattle, USA. e-mails: popejoy@uw.edu; smfullrtn@uw.edu

1. Need, A. C. & Goldstein, D. B. *Trends Genet.* **25**, 489–494 (2009).
2. Bustamante, C. D., De La Vega, F. M. & Burchard, E. G. *Nature* **475**, 163–165 (2011).
3. Fullerton, S. M., Yu, J.-H., Crouch, J., Fryer-Edwards, K. & Burke, W. *Hum. Genet.* **127**, 563–572 (2010).
4. Carlson, C. S. et al. *PLoS Biol.* **11**, e1001661 (2013).
5. Desta, Z., Ward, B. A., Soukhova, N. V. & Flockhart, D. A. *J. Pharmacol. Exp. Ther.* **310**, 1062–1075 (2004).
6. Daly, A. K. *Nature Rev. Genet.* **11**, 241–246 (2010).
7. Fumagalli, M. et al. *Science* **349**, 1343–1347 (2015).
8. Lek, M. et al. *Nature* **536**, 285–291 (2016).
9. Petrovski, S. & Goldstein, D. B. *Genome Biol.* **17**, 157 (2016).
10. Manrai, A. K. et al. *N. Engl. J. Med.* **375**, 655–665 (2016).
11. Li, Y. R. & Keating, B. J. *Genome Med.* **6**, 91 (2014).
12. Yancey, A. K., Ortega, A. N. & Kumanyika, S. K. *Annu. Rev. Public Health* **27**, 1–28 (2006).
13. Leboy, P. S. & Madden, J. F. *DNA Cell Biol.* **31**, 1365–1371 (2012).





Washington DC at night.

# Scientists must have a say in the future of cities

A United Nations conference seeks urban sustainability. But the agenda will fail without input from researchers, warn **Timon McPhearson** and colleagues.

**M**ore urban areas will be built in the next 30 years than ever before. Growing settlements will increase demand for infrastructure, food, energy, water and housing. Simply meeting the projected urban expansion will breach the warming limit set by the 2015 Paris climate agreement.

This week, the United Nations' third major global cities conference, Habitat III, convenes in Quito, Ecuador. Held every 20 years, this multilateral meeting will adopt a global framework for making cities more sustainable — the New Urban Agenda (NUA). Sadly, science was largely absent from the drafting process of the NUA. By contrast, expert evidence guided the Paris climate deal, the 2015 Sendai Framework

for Disaster Risk Reduction and the UN's 2030 Agenda for Sustainable Development and its Sustainable Development Goals (SDGs).

One reason is that the scientific community was unprepared for Habitat III. The few scientists invited to participate accepted a consultative role, nested among other public voices. Then, in late July, negotiators dropped the proposed multistakeholder panel, which would have formally embedded scientists and other non-state representatives in the implementation process. European Union members and other rich countries were concerned that the panel would be expensive. The final draft of the NUA<sup>1</sup> brokered in New York last month failed to reverse this. It is thus necessary to argue the

case once again for the importance of urban science and of establishing a science-policy interface for the NUA.

Urban research is disparate, marginalized and ill-prepared to interact effectively with global policy. The Habitat III agenda requires a global community of urban biophysical and social scientists to assess developments and help direct progress. To achieve the SDGs and the NUA, the global urban research community must come together to develop institutions, funding mechanisms and research agendas.

## URBAN ACCELERATION

Rapid urbanization is one of the biggest social transformations in human history<sup>2</sup>. Cities are depleting resources and face new



risks caused by climate change. For example, disastrous floods in the past decade in the United States, Philippines, the United Kingdom, India and China show how vulnerable coastal and riverside cities are to storm surges, with trillions of dollars' worth of assets at stake<sup>3</sup>.

Yet cities can also be engines of innovation. Here, the most progress is being made on climate change<sup>4</sup> and other sustainability goals<sup>5</sup>. For example, cities around the world are embracing nature-based infrastructure for adaptation and resilience, such as green roofs and wetland restoration<sup>6</sup>.

City processes are complex and often far from equilibrium, displaying emergent properties and non-linear dynamics. Urban areas are difficult to plan, manage and govern, and have a rapacious appetite for energy and materials, with global environmental impact<sup>7</sup>. Urban challenges ask complex and interrelated questions about equity, justice, resilience, economic opportunity, infrastructure development, ecological restoration and more.

**"Most research is in the north; most need is in the south."**

## COME TOGETHER

Implementing, monitoring, evaluating and revising the NUA and related SDGs will require evidence from across the research community, from natural and social scientists to humanities scholars. To be useful to policymakers, urban research needs to be organized, representative and seen as legitimate. This is far from the case.

Urban researchers are scattered across non-governmental organizations, government agencies and community-based organizations, and are found both in and outside academia. They span many disciplines and professions, including architecture, ecology, engineering and geography. People, funds and institutions are distributed unevenly.

Most urban scientists and resources are located in the global north and in large cities, but the most pressing urban challenges tend to be found in the global south and in small to medium sized cities. Urban research and solutions are context-specific. The different developmental trajectories of cities in Africa, Asia or Latin America may be at least as significant as the better-documented gap between northern and southern cities<sup>8</sup>.

Scholars must expand primary research in less-studied and rapidly changing urban contexts such as those found in Latin America and south and southeast Asia. For example, too little is known about the global interlocking system of cities in terms of material usage, ecosystems, social and political norms, migration, disease vectors

and innovation. Urban scientists need to better map and model these to provide information for planning, management and policymaking<sup>9</sup>.

The skills of drawing together many sources of knowledge to inform global urban policy are in short supply. Professional certification systems and the lack of interactions across sectors reinforce the isolation of specialists such as engineers, architects and planners. Paradoxically, urban researchers in the global south, who are forced to become generalists because of skill shortages, may have broader experience than their peers in the north, where academic practices and evaluations often reinforce specialization. Many southern scholars engage directly with urban communities and local and national policymakers.

We view as inadequate one model that has been advocated during the Habitat III process for bringing together urban knowledge at the global scale — an urban equivalent of the Intergovernmental Panel on Climate Change (IPCC). Although the IPCC has been successful in focusing the efforts of the international climate-science community around specific policy-relevant questions, it has also proved slow and cumbersome. Urban science is broad and fast-moving. Even urban scientists do not necessarily agree on the most important research questions, let alone the prescriptions.

## FIVE STEPS

We advocate the following steps to boost the development and impact of urban science.

**Form a global urban scientific body.** An international urban science platform should be formed to address the post-2030 agenda. It must enable broad science-policy interaction and cross-city learning at a global scale. This could connect existing global networks such as the IPCC, Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services, UN-Habitat, the UN Environment Programme, the Future Earth Urban Knowledge-Action Network and the Sustainable Development Solutions Network. The body should be developed in consultation with scientists, professional societies and holders of urban knowledge at all levels, including scholars, civil servants and citizens. Governance must be inclusive and could be based on the polycentric model developed at Future Earth, a global platform for sustainability research, with regional hubs responding to local issues.

**Spread knowledge and institutions globally.** Most research is in the north; most need is in the south. Inclusivity and diversity across geographic regions

and scientific domains is key to legitimacy and legibility. Major investment is needed in academic institutes that are sited at the nexus of urban research, policy and practice in rapidly urbanizing cities. Mapping knowledge and institutions would help to uncover key geographic and thematic gaps.

**Boost funding for urban research.** Truly global sources of research grants are needed to allow cross-comparison studies of cities and regions. These should be set up with support from national governments, development banks and private foundations. This would require large sums (one of the reasons that the multistakeholder panel was taken out of the final NUA draft).

**Support transdisciplinary research and synthesis.** Communities with relevant knowledge must guide urban-development policy over the short and long term. Transdisciplinary research must be supported through new sources of urban science funding and organizations. Existing knowledge should be synthesized and fed into policymaking at all levels.

**Improve access to science-policy arenas.** Urban scholars must have a clear role in the policy platforms that are emerging in the NUA and the wider multilateral system, such as the links forming between the urban SDGs and the Future Earth Urban Knowledge-Action Network.

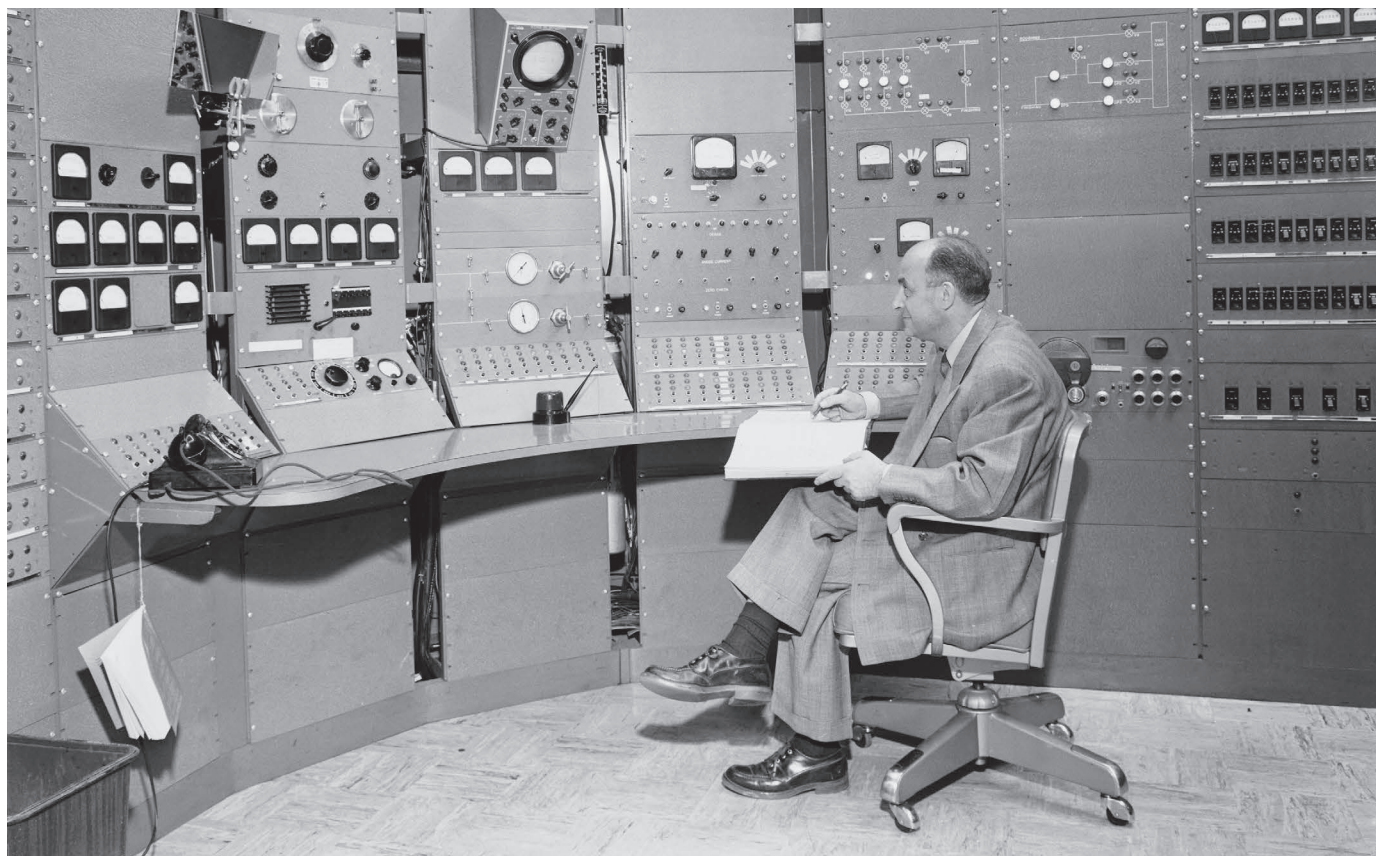
It is imperative to scale up urban research and foster a scientific leadership to direct and critique global urban policymaking. ■

**Timon McPhearson** is a visiting research scientist at the Cary Institute for Ecosystem Studies and associate professor of urban ecology at The New School in New York, USA. **Susan Parnell, David Simon, Owen Gaffney, Thomas Elmqvist, Xuemei Bai, Debra Roberts and Aromar Revi.**  
e-mail: [timon.mcphearson@newschool.edu](mailto:timon.mcphearson@newschool.edu)

1. *Habitat III: Draft New Urban Agenda* (United Nations, 2016).
2. Bai, X., Shi, P. & Liu, Y. *Nature* **509**, 158–160 (2014).
3. Aerts, J. C. J. H., Botzen, W. J. W., Emanuel, K., Lin, N., de Moel, H. & Michel-Kerjan, E. O. *Science* **344**, 473–475 (2014).
4. Revi, A. & Rosenzweig, C. *The Urban Opportunity: Enabling Transformative and Sustainable Development* (Sustainable Development Solutions Network, 2013).
5. Kanuri, C., Revi, A., Espey, J. & Kuhle, H. *Getting started with the SDGs in Cities* (2016); available at: <https://sdgcities.guide>
6. Roberts, D. et al. *Environ. Urban.* **24**, 167–195 (2012).
7. Grimm, N. B. et al. *Science* **319**, 756–760 (2008).
8. Parnell, S. & Oldfield, S. *The Routledge Handbook on Cities of the Global South* (Routledge, 2014).
9. McPhearson, T. et al. *BioScience* **66**, 198–212 (2016).

A full list of author affiliations accompanies this article online: see [go.nature.com/2deyrym](http://go.nature.com/2deyrym).





BETTMAN/GETTY

Enrico Fermi at the controls of the synchrocyclotron particle accelerator at the University of Chicago, Illinois, in the 1950s.

## PHYSICS

# Fallible pontiff of physics

**Graham Farmelo** assesses a biography of star theorist–experimentalist Enrico Fermi.

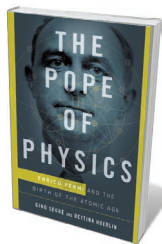
The megalomaniac physicist Edward Teller “was always certain that he was smarter than all his colleagues on the Manhattan Project, except one”, a younger colleague of his (Robert Sachs) told me in 1993. That exception was Enrico Fermi, “the nearest we physicists had to a Pope”.

Fermi’s infallibility is one of the prominent themes of *The Pope of Physics*, the first popular cradle-to-grave biography in English of the most famous Italian scientific investigator since Galileo Galilei. The authors are married: Bettina Hoerlin, formerly a health executive, and Gino Segrè, a nephew of one of Fermi’s most distinguished colleagues, Emilio Segrè. They each met their subject independently, and plainly found his career an inspiration.

Fermi’s many contributions to nuclear science and technology — such as the creation of the first nuclear reactor — have been recounted ad nauseam, so he might not seem the most promising candidate for a biography. Segrè and Hoerlin, however, seek a new perspective. They quickly hit their stride with

a lucid account of how Fermi was born in 1901 to a middle-class family in Rome and became one of the very few physicists to be in the front rank in both theory and experiment. The authors engagingly describe how Fermi taught himself basic mathematics and physics to a high standard by the time he was 17. Blessed with a prodigious ability to identify the essence of every physics problem, he matured as a researcher in the 1920s, at just the right time to make his mark on two open frontiers: nuclear science and quantum theory.

Fermi was often quicker than more formal thinkers to apply abstract ideas to improve the understanding of nature. In 1926 he became



**The Pope of Physics: Enrico Fermi and the Birth of the Atomic Age**  
GINO SEGRÈ AND BETTINA HOERLIN  
Henry Holt: 2016.

the first to use quantum theory to study large aggregations of electrons, one of a class of subatomic particles later called fermions, and he later set out the first quantum field theory of radioactive  $\beta$  decay. Yet his forte was experimental physics, as he demonstrated at the Sapienza University of Rome, which appointed him a professor at the age of 24. He made the Italian capital one of the world’s most productive centres of modern physics.

Fermi and his talented colleagues found that slow neutrons are remarkably effective at inducing radioactivity in some heavy chemical elements. That won him the physics Nobel in 1938. However, when Fermi and his group studied the products of some of the nuclear reactions that they induced, their interpretations were sometimes wrong, as he later acknowledged. To be fair, the great radiochemist Otto Hahn and his group in Berlin made the same mistake at about the same time, before they recognized that uranium nuclei undergo the process later called fission.

The authors describe movingly how the rise of Fascism in Italy led the Fermi family to emigrate to the United States. Fermi arrived in New York City on 2 January 1939, and soon took up a post at Columbia University. Although some US officials were wary of putting foreigners in leading positions in secret projects, Fermi was indisputably the best person to lead the development and construction of the first nuclear reactor, CP-1, which began operation on 2 December 1942: a crucial stage in the development of nuclear weapons. Robert Oppenheimer invited him to join the Manhattan Project and created a unit named after him. 'F division' dealt with especially tough problems in experimental and theoretical nuclear physics.

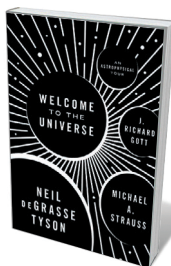
The biography sometimes has a hagiographic whiff. Segrè and Hoerlin agree with the consensus that he was largely apolitical and was "a scientist pure and simple", as his Hungarian colleague Leo Szilard once described him. I suspect that the truth is deeper, although it will probably remain hidden. The authors tiptoe around some of his less endearing characteristics, such as his ability as a father. Fermi's daughter told me in 1992 that he was "always a distant figure", and that 40 years earlier her father had approached her out of the blue one day, suggesting that they get know each other better by doing a few experiments around the house on the then-new material Silly Putty. He did not repeat this initiative, which did little to improve their relationship.

In my view, Segrè and Hoerlin underplay Fermi's considerable influence on young physicists after the Second World War. It would have been revealing to read, for example, more from the great theoreticians Murray Gell-Mann and Chen-Ning Yang, who worked closely with him. Likewise, it would have been rewarding to have heard more about Fermi's ideas on the origin of cosmic rays and his thinking about the future of subatomic-particle accelerators. In January 1954, he gave a far-sighted lecture in which he envisaged the possibility of building an ultrahigh particle accelerator that girdled the entire planet by 1994.

Ten months later, he was dead. Lying in his hospital bed the day after he had learnt that he had terminal stomach cancer, he told the astrophysicist Subrahmanyan Chandrasekhar: "For a man past 50, nothing essentially new can happen." Had he lived two decades longer, that would not have been true. Physicists were soon to make discoveries that fundamentally altered their understanding of nature — another reminder that Fermi's foresight was, like that of all putative scientific popes, fallible. ■

**Graham Farmelo** is a by-fellow at Churchill College in Cambridge, UK.  
e-mail: [graham@grahamfarmelo.com](mailto:graham@grahamfarmelo.com)

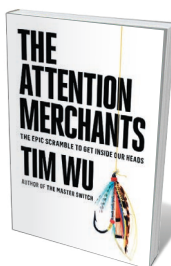
## Books in brief



### Welcome to the Universe: An Astrophysical Tour

Neil deGrasse Tyson, Michael A. Strauss and J. Richard Gott  
PRINCETON UNIVERSITY PRESS (2016)

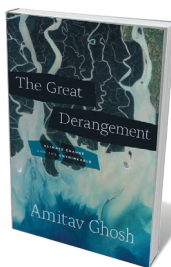
As citizens of the cosmos, we are duty bound to explore it. So opine astrophysicists Neil deGrasse Tyson, Michael Strauss and Richard Gott, guides on this bracing expedition through dusty galactic hinterlands and the vast theoretical vistas of Albert Einstein's work. Each is a master at untangling the abstruse through metaphor: Tyson crams 100 million elephants into a thimble to illustrate neutron-star density, and Gott recounts John Archibald Wheeler demonstrating entropy by mixing tea and water and throwing it into a 'black hole'.



### The Attention Merchants: The Epic Scramble to Get Inside Our Heads

Tim Wu KNOFF (2016)

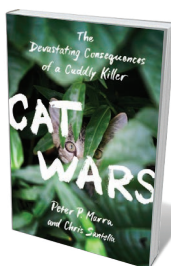
Media scholar Tim Wu plunges into the noisome history of "attention harvesting" — the commodification of human attention by industry and government. It began, Wu reveals, with the juxtaposition of advertisements and lurid news in 1830s gutter journalism, and persisted in the engineered demands of "scientific advertising", the efforts of propagandist Edward Bernays (who persuaded women to smoke) and the infiltration of fast-food ads into US schools. To evade this induced narcosis and reclaim lived experience, Wu argues, we must wean ourselves off the digital.



### The Great Derangement: Climate Change and the Unthinkable

Amitav Ghosh UNIVERSITY OF CHICAGO PRESS (2016)

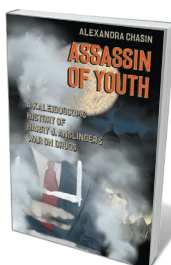
Resistance to the grim realities of climate change is so widespread that the crisis barely figures in literary fiction, notes writer Amitav Ghosh. Branding our era of denial and inertia the Great Derangement, Ghosh looks in turn at literature, history and politics to examine this failure, noting that extreme events such as 2012's Hurricane Sandy are so freakish that they seem inexpressible. The solution, he argues, lies in collective action as well as scientific and governmental involvement — and in a resurgence in our imaginative capacity to envision human existence anew.



### Cat Wars: The Devastating Consequences of a Cuddly Killer

Peter P. Marra and Chris Santella PRINCETON UNIVERSITY PRESS (2016)

Among the hundreds of millions of domestic cats, many range freely. That group is effectively a death squad for songbirds, killing an estimated 4 billion US avifauna a year; globally, island cats drive 14% of vertebrate extinctions. This deeply researched overview by conservation scientist Peter Marra and writer Chris Santella interlaces discussions of feline domestication and avian conservation with the science of decline and of feline spillover diseases. It culminates with a stark choice: control free-ranging cats or witness the ongoing erosion of affected ecosystems.



### Assassin of Youth: A Kaleidoscopic History of Harry J. Anslinger's War on Drugs

Alexandra Chasin UNIVERSITY OF CHICAGO PRESS (2016)

Harry Anslinger helmed the US Federal Bureau of Narcotics from 1930 to 1962, shaping US drug policy through what Alexandra Chasin calls "an elaborately disastrous set of policies and laws". In this idiosyncratic chronicle, Chasin paces the trail from temperance to today, when nearly half the inmates of US jails are incarcerated for drug offences. A sorry tale of how one man's racial prejudice and predilection for prohibition led to a colossal policy failure. **Barbara Kiser**



## HISTORY

# Einstein the statesman

Nancy Thorndike Greenspan enjoys a study of the physicist as engaged public figure.

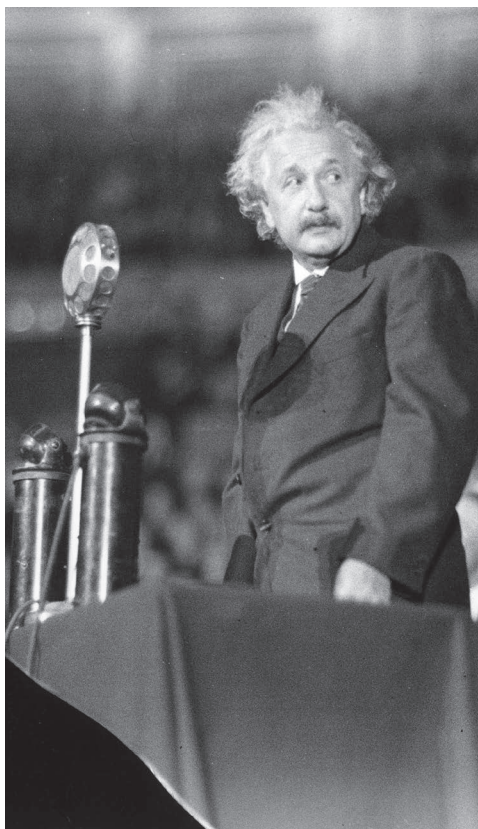
Albert Einstein's dreams have been analysed, his love life scrutinized, his letters parsed. His brain has been dissected and driven 5,000 kilometres across the United States. In a way, he has survived it all.

In *Einstein and Twentieth-Century Politics*, it is the physicist's politics and principles that are once again debated. Einstein is the global thinker, a scientist with important social insights and a charismatic personality besides. Richard Crockatt shows that the causes that he championed embody the world's conflicts of the time: pacifism; Zionism and Israel; the atomic bomb and the arms race; and world government. Writes Crockatt (a British scholar of US foreign policy and mid-twentieth-century international relations): "In looking at his life and ideas one sees over his shoulder into the world he lived in, not merely into his mind."

Einstein's first public foray into politics was in 1914. He was bold from the start: aged 35 and teaching at the University of Berlin, he was one of only four signatories to a manifesto against the First World War. It was to counter another signed by 93 of his peers, defending Germany's 'just' war. He didn't state so openly, but Einstein hoped that Imperial Germany would lose the war and rise from the ashes as a democratic socialist government.

Before the war, he was little-known outside physics. Celebrity enveloped him in 1919, when experimentalist Arthur Eddington confirmed the general theory of relativity. After this, Einstein — who was Jewish — was no longer a mere mortal, a fact that roused an increasingly anti-Semitic Germany. Anti-relativists with an anti-Jewish agenda attacked him. In this and only this instance, he lashed out in a newspaper article, which brought more attacks. He learned that detractors could agitate him, that he needed to protect his privacy and that entering the political fray disrupted his science. From then on, he made pronouncements on political and moral issues that aimed to keep him above ordinary politics. These dimensions are all nimbly woven into *Einstein and Twentieth-Century Politics*.

After the First World War, the League of Nations was founded to advance peace. Einstein had an ambivalent relationship with it, as did many of his liberal contemporaries. They feared that the military strength of nations would thwart their pacifist goal. The devastation of the war had made many a pacifist purist. But in 1933, having left Germany



Einstein speaking on science and civilization in 1933.

for the United States in the face of the Nazi assault, Einstein refined his views. A grudging pragmatism took hold: what good would come from supporting peace if freedom were lost? Allied military strength became a benefit. Such was his evolution on many issues, garnering criticism in private from his leftist intellectual friends and in public from his foes, especially those in the US government. He was characterized variously as reckless, hypocritical, naive and pro-Soviet.

One decision — promoting the building of the atomic bomb — he regretted until he died. He wrote three letters to US President Franklin Roosevelt: the first two advocated for the bomb; the last indirectly warned against using it. After the Second World War, he explained his rationale

— and that he never worked on the bomb — but he could not break the public's perception that he was responsible for it. In the end, all he could do was to press more urgently for a supranational government that controlled military power, especially the bomb.

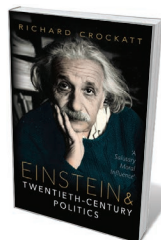
Crockatt skilfully uses letters and articles from "the liberal wing of international opinion", non-scientists all, to limn the subtleties of the issues. Philosophers Bertrand Russell and Albert Schweitzer; writers George Bernard Shaw, Thomas Mann and H. G. Wells; and Mahatma Gandhi — with a cameo by Sigmund Freud — all draw out the nuances in Einstein's positions and expose the complexities created by a rapidly changing world. They agreed, more or less, on the objective: a world government that limits the power of the nation-state and nullifies war. Details, such as the role of the state, create the debate.

Crockatt's deft pen and thoughtful approach form an engaging and revealing image of Einstein the non-scientist. He marshals a wealth of material into a convincing narrative. However, to tease out Einstein's "distinctive intellectual and emotional qualities", Crockatt may have overreached. In essence, he alleges that Einstein examined every angle of a scientific idea but "rarely, if ever, questioned" his political principles because they were self-evident.

I doubt it was that black and white. Crockatt himself describes numerous incidents in which Einstein altered his political positions fundamentally. Crockatt suggests that the shifts were tactical, but they must have required deep reflection and questioning. These were profoundly consequential decisions. Nor was Einstein entirely the dispassionate, empirical scientist of his own characterization or that Crockatt assumes. Einstein's argument for rejecting quantum mechanics, that God "does not play dice", was more about intuition than well-founded scientific principle.

Nevertheless, *Einstein and Twentieth-Century Politics* delivers what Crockatt promises: a picture of Einstein's world and mind, conveyed with an insightful brush. And these issues political and philosophical are sufficiently complex that they will surely live to see many other interpretations. ■

Nancy Thorndike Greenspan is a writer based in Washington DC. Her most recent book is *The End of the Certain World: The Life and Science of Max Born*. e-mail: [nancy@nancygreenspan.com](mailto:nancy@nancygreenspan.com)



**Einstein and Twentieth-Century Politics: 'A Salutary Moral Influence'**  
RICHARD CROCKATT  
Oxford University Press: 2016.

# Correspondence

## Don't let witch hunts taint investigations

Donald Kornfeld and Sandra Titus argue that misconduct should be considered when investigating irreproducible research (*Nature* **537**, 29–30; 2016). In my view, this premise of 'guilty until proved innocent' risks turning a scholarly investigation into a witch hunt.

Distinguishing poorly designed research from faked data is hard, but it is generally more difficult to prove misconduct than to identify the cause of irreproducibility. Moreover, investigating bad science costs less than examining misconduct (in terms of money, time, careers and so on). One estimate put the direct cost of a misconduct enquiry at US\$525,000 (A. M. Michalek *et al.* *PLoS Med.* **7**, e1000318; 2010).

The authors assert that the US National Institutes of Health's (NIH) training mandate for responsible conduct of research failed to reduce misconduct. Perusing research on the rate of misconduct, I find estimates that span several orders of magnitude, so it is unclear whether misconduct is or has been rising or falling. Thus, we cannot say what effect the NIH training scheme has had.

The damage caused to the scientific record by publishing sloppy, plagiarized or fabricated research is ultimately the same.

**Kenneth Pimple** *Indiana University, Bloomington, USA.*  
pimple@indiana.edu

## University on the rise without PhD students

California State University in Northridge (CSUN) is ranked 24th in the latest *Nature Index* of the 25 North American institutions classed as 'Rising stars' (see [go.nature.com/2dfvirb](http://go.nature.com/2dfvirb)). Of these, CSUN has the greatest percentage increase in publications in 2012–15 in

high-impact journals (up by 190.61%). As professor emeritus at CSUN, my view is that the rise is attributable to the university's unusual research model.

The university does not offer any PhD degrees. Instead, science undergraduates and master's students work alongside exceptional faculty members, who train and mentor them in research. The faculty includes 28 prestigious members hired over the past decade. The dean of the science college, Jerry Stinner, backed by the university chairs, president and provost, organized CSUN funding to recruit and support them. This support included comprehensive packages of research equipment and supplies, specific laboratory renovations and a reduced teaching load while they set up their labs.

**Steve Oppenheimer** *California State University, Northridge, California, USA.*  
steven.oppenheimer@csun.edu

## Hasten end of dated fossil-fuel subsidies

Nations at last month's G20 summit in China reaffirmed their 2009 commitment to phase out fossil-fuel subsidies, echoing a call from almost two decades ago to end subsidies that are "adverse in the long run to both the economy and the environment" (N. Myers *Nature* **392**, 327–328; 1998).

Similar 'perverse' subsidies continue to encourage logging of the few remaining pockets of old-growth forest in western Canada and overfishing in the high seas. Yet the fossil-fuel industry receives the largest subsidy of all, estimated by the International Monetary Fund (IMF) last year at US\$1,000 annually for every citizen in the G20 group. Most of this is provided by countries with energy taxes that are too low to cover the adverse effects of fossil-fuel consumption on human health and the environment ([go.nature.com/2dbz2zf](http://go.nature.com/2dbz2zf)).

The IMF also estimates that eliminating fossil-fuel subsidies would cut global carbon dioxide emissions by more than 20% and raise government revenues by \$2.9 trillion (or 3.6% of global gross domestic product). Such a step would save up to \$93 per tonne of greenhouse-gas emissions removed (see [go.nature.com/2dowcw](http://go.nature.com/2dowcw)).

These sums alone would fund climate adaptation and the protection of imperilled global biodiversity for the next 30 years (D. P. McCarthy *et al.* *Science* **338**, 946–949; 2012). The money would also boost development of renewable energy sources and domestic support for a low-carbon economy.

**Tara Martin** *University of British Columbia, Vancouver, Canada.*  
tara.martin@ubc.ca

## Are farmed fish just for the wealthy?

Christopher Golden and colleagues argue that farmed fish contribute little to global food security because they are "mostly exported to the wealthy countries of Europe and North America" (*Nature* **534**, 317–320; 2016). In fact, more than 90% of farmed fish produced in China, India, Indonesia, Bangladesh, Egypt, the Philippines and Myanmar — some of the world's largest aquaculture-producing developing countries — remains in domestic markets (see [go.nature.com/2dqwwh](http://go.nature.com/2dqwwh)).

Aquaculture products are more accessible to the poor in many developing nations than ever before (K. A. Toufique and B. Belton *World Dev.* **64**, 609–620; 2014). And the aquaculture boom of the past two decades has stabilized world fish prices (S. Tveterås *et al.* *PLoS ONE* **7**, e36731; 2012).

The realities of the supply and demand of aquaculture products mean that these now complement capture fisheries for global food and nutrition security.

**Ben Belton** *Michigan State*

*University, East Lansing, USA.*  
**Simon R. Bush** *Wageningen University, the Netherlands.*  
**David C. Little** *University of Stirling, UK.*  
beltonbe@msu.edu

*Christopher Golden et al. reply* — Our argument is that most farmed fish are not reaching nutritionally vulnerable people in the low-income, food-deficit countries of sub-Saharan Africa and the Pacific islands (*Nature* **534**, 317–320; 2016). In those nations, fish is a traditional food source that comes primarily from capture fisheries, including subsistence harvests (M. M. Dey *et al.* *Mar. Policy* **67**, 156–163; 2016). Domestic consumption and import of aquaculture products are still relatively insignificant (see [go.nature.com/2dinzuc](http://go.nature.com/2dinzuc)).

In such places, aquaculture policy interventions need to be optimized for nutritional value and distribution to food-insecure populations. This could be achieved through appropriate regulations and market instruments (such as tax incentives or subsidies) and public-health campaigns, in close alliance with conservation strategies for sustainable fisheries.

**Harvard T. H. Chan School of Public Health, Cambridge, Massachusetts, USA.  
golden@hsph.harvard.edu**

### CORRECTIONS

The Outlook article 'Industrial strength' (*Nature* **537**, S57–S59; 2016) incorrectly stated that the 1999 trial at the University of Pennsylvania was based on a retrovirus; it was in fact based on an adenovirus.

Also, the Outlook Q&A 'Illuminated Universe' (*Nature* **537**, S205; 2016) incorrectly gave the amount of dark energy in a cubic metre of space as 10–27 kilograms instead of 10<sup>–27</sup> kilograms.



# Roger Yonchien Tsien

## (1952–2016)

Creator of a rainbow of fluorescent probes that lit up biology.

Roger Yonchien Tsien pioneered the use of light and colour to ‘peek and poke’ at living cells to see how they work. His most famous achievement, recognized by a share of the Nobel Prize in Chemistry in 2008, transformed biology: he developed a rainbow of probes, based on the jellyfish green fluorescent protein (GFP), to illuminate cell structure and function.

Roger died suddenly in a park near his home in Oregon on 24 August. He was born in New York in 1952 with science in his blood. His father’s cousin was Tsien Hsueshen (Qian Xuesen), architect of China’s missile and space programme. Roger would combine his father’s engineering talent with the medical interests of his mother, a nurse.

Roger had an early passion for chemistry. Despite his Chinese name (which means ‘always healthy’), childhood asthma often kept him indoors, reading and drawing. He fought going to kindergarten until his teacher allowed him to bring in a favourite book: he picked *All about the Wonders of Chemistry*. From the age of eight, he performed increasingly complex and sometimes hazardous chemistry experiments at home. At 16, he went to Harvard University in Cambridge, Massachusetts (avoiding the Massachusetts Institute of Technology, where his father, uncles and brothers studied), and sampled many subjects. Ironically he found the chemistry courses “so distasteful” that he abandoned them for neurobiology.

Roger then spent nine years at the Physiological Laboratory at the University of Cambridge, UK. First he was a PhD student with the eminent muscle physiologist Richard Adrian; then he did a postdoc with one of us (T.J.R.). He emerged as an ingenious, largely self-taught synthetic chemist.

Much of Roger’s early work was directed at imaging neural activity, by trying to develop tracers of sodium- or calcium-ion movements that support brain signalling. By 1980, he had invented quin2, a synthetic fluorescent dye that selectively binds to calcium, and had devised a clever way to sneak this dye and other probes into intact cells. This first practical probe for calcium found wide early use in studies of intracellular calcium signalling.

Amazingly, Roger struggled to find a faculty position because his work straddled disciplines. In 1982, he joined the physiology department at the University of California,



Berkeley, where colleagues encouraged him to create more tools. First came superior calcium dyes, in particular fura2, which is strongly excited by different wavelengths of ultraviolet light before and after binding calcium. Capitalizing on this feature of fura2 (and indicators with similar optical properties), Roger and his group made it much easier to monitor calcium under challenging conditions, for example, across the width of a cell. His group also created valuable fluorescent sensors for pH and for sodium.

In 1989, facing resource constraints, Roger transferred to the University of California, San Diego (UCSD). Here he remained for the rest of his career. He wanted to make sensors that could be genetically encoded, allowing researchers to target specific cell types without having to inject a tracer. In the 1990s, he saw the potential of GFP. The protein had been isolated from jellyfish in the 1960s by Osamu Shimomura (who shared the 2008 Nobel) and cloned by Douglas Prasher in 1992. Martin Chalfie, who also shared in the Nobel, first used GFP to image living cells in 1994.

Roger’s lab pioneered the development of GFP variants. Through a combination of rational design and random mutagenesis, they created dozens of bright fluorescent proteins of various colours based on GFP. Roger later produced longer-wavelength sensors based on red fluorescent proteins.

He took great pleasure in naming probes after fruits such as the tomato, cherry and plum.

GFP variants are now ubiquitous in biological research. They can be used to bind with and track cancer cells, aid gene therapy, image mitosis, paint neurons in rainbow colours and spy on signalling in subcellular organelles such as mitochondria. They have even been used to make art.

Roger’s group at UCSD developed many other optical probes, including fast-response sensors to measure electrical signals across cell membranes, and dyes for tracking proteins with a combination of light and electron microscopy. In recent years, he had two main projects: the design of fluorescent tracers to illuminate tumours during cancer surgery; and the storage of long-term memory by the pattern of holes in the perineuronal net that surrounds neurons in the brain.

Roger’s trajectory helped to make it respectable, indeed fashionable, to spend a career inventing reagents and methods. He is named in more than 160 US patents, often as lead inventor. Although naturally keen to participate in the first application of his new tools, he was also generous in providing materials to other scientists.

Roger co-founded three biotech companies that capitalized on his inventions. He semi-seriously quipped to his wife Wendy that, apart from the potential human benefit, the main point of these companies was to provide suitable jobs for his postdocs.

Roger was a fine pianist and briefly considered a musical career. A gifted amateur photographer — a hobby in keeping with his passion for colour and imaging — he enjoyed holidays in the wild outdoors, often taking arduous treks, camera in hand.

Roger will be hugely missed by family, friends, colleagues and the many scientists who appreciated him as a brilliant enabler of scientific progress. ■

**Timothy J. Rink** is an independent biotechnology consultant. He worked with Roger Tsien at Cambridge. **Louis Y. Tsien** is Roger’s elder brother and lives in Watertown, Massachusetts. **Richard W. Tsien** is Roger’s eldest brother and a neuroscientist at New York University, New York, USA.  
e-mails: rinktj@gmail.com;  
louis.tsien@gmail.com;  
richard.tsien@nyumc.org

HOLGER MOTZKAU/WIKIPEDIA/WIKIMEDIA COMMONS

## EARTH SCIENCE

## Megafloods downsized

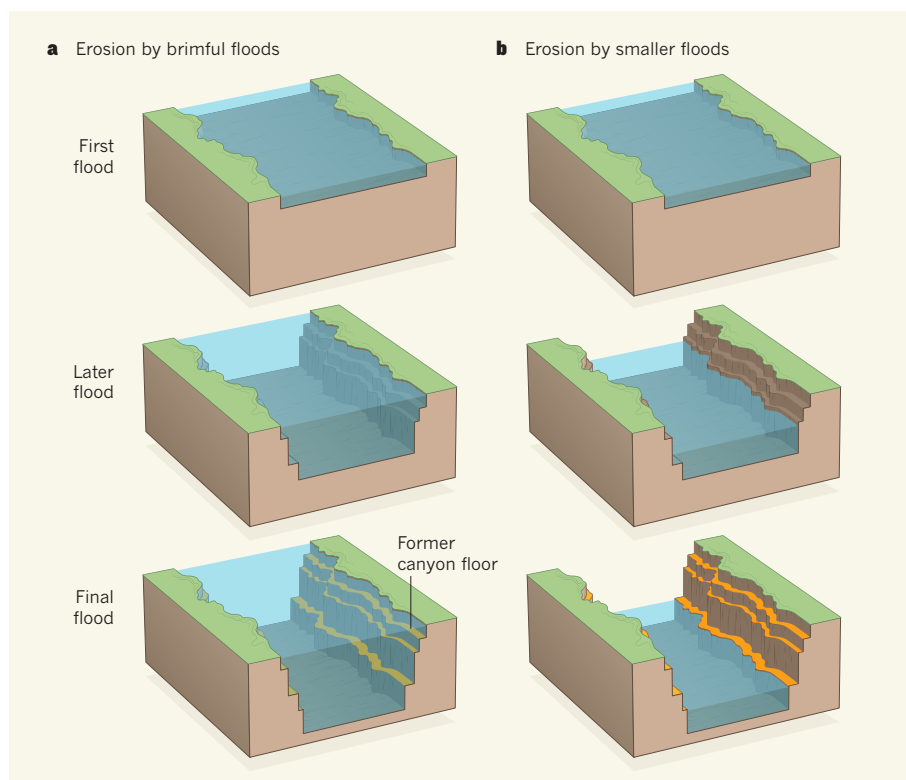
A fresh look at the Channeled Scablands of North America shows that the ancient floods that scarred that landscape were smaller than is commonly assumed. This result could revise estimates of similar floods on Mars. [SEE LETTER P.229](#)

J. TAYLOR PERRON & JEREMY G. VENDITTI

The enormous canyons of the Channeled Scablands in the northwestern United States, many of which contain no rivers, puzzled geologists for decades. The gradual realization that these canyons were carved thousands of years ago by huge floods spawned by melting glaciers challenged the idea that Earth's surface is shaped by gradual, steady erosion. However, on page 229, Larsen and Lamb<sup>1</sup> show that at least one of the canyons was formed by a succession of much smaller floods, a finding that has implications for flood-carved canyons on Mars.

When the geologist J Harlan Bretz proposed in the 1920s that the Channeled Scablands were created by a catastrophic flood<sup>2</sup>, his ideas were attacked relentlessly by geologists who subscribed to the mainstream view that erosion is slow and steady, and who wanted to distance their profession from the notion of a biblical deluge. Bretz did not identify the source of the flooding until the 1940s, when his colleague Joseph Pardee found evidence<sup>3</sup> that ancient Lake Missoula, which formed at the margin of the melting Cordilleran ice sheet roughly 15,000 years ago, had drained catastrophically to the west. This discovery led to the gradual acceptance of Bretz's flood hypothesis, which was later supported by studies that considered the mechanics of large flows through canyons<sup>4</sup>. Subsequent analyses of sediments deposited throughout the region showed that the Channeled Scablands had experienced not one but many floods<sup>5</sup>.

Although the flood origin of the Channeled Scablands is no longer disputed, the sizes of the individual floods remain uncertain. It has become common practice to place an upper bound on the flow rate of the floods by assuming that they filled the present-day canyons to the brim. Estimated flood magnitudes based on this assumption<sup>6</sup> range up to 60 cubic kilometres per hour — nearly 100 times the average flow rate of the Amazon River today<sup>7</sup>. But these estimates might be much too large. Glaciologists have argued that it is difficult for ice sheets to store enough water to produce such enormous floods<sup>8</sup>. The brimful-flood model also requires the unlikely scenario that each flood passing through the canyons was



**Figure 1 | Competing models of canyon erosion by floods.** **a**, It is commonly assumed that canyons form in accordance with a brimful model, which requires progressively larger and deeper floods as the canyon erodes. **b**, Larsen and Lamb<sup>1</sup> use remnants of former canyon floors to show that Moses Coulee was instead shaped by a sequence of smaller floods.

larger than the one that preceded it, because the canyon deepens as each successive flood erodes the bedrock (Fig. 1a).

Larsen and Lamb waded into this debate and present evidence that a series of consistently sized, moderate floods eroded the canyons of the Channeled Scablands. In this scenario, the first flood filled the shallow, newly formed canyons to the brim, but subsequent floods only partly filled the deepening canyons (Fig. 1b). They studied Moses Coulee (Fig. 2), a canyon in which a series of bench-shaped terraces preserves the remnants of former canyon floors that were abandoned by the flood water as the canyon was progressively eroded.

Using previous estimates of the forces required to erode blocks of rock from the canyon floor, and a computational model of flood flow through the canyon, the

authors constrained the minimum flow rate corresponding to each remnant canyon floor. Their calculated flow rates are consistent with the presence of gravel bars that the most recent floods deposited in the canyon. Brimful floods would have instead suspended the gravel (and even larger boulders) high in the flow, preventing deposition. Larsen and Lamb conclude that Moses Coulee was eroded by repeated floods of no more than  $2 \text{ km}^3 \text{ h}^{-1}$ . This flow rate is by no means small — it is more than three times that of the Amazon River<sup>7</sup> — but it is much smaller than the maximum of  $10 \text{ km}^3 \text{ h}^{-1}$  that is implied by the brimful model for Moses Coulee.

Floods as large as those discussed by Larsen and Lamb have not been observed in recorded history. This makes it difficult to test some of the authors' assumptions, such as the estimated





**Figure 2 | Moses Coulee.** Upstream view along the east wall of Moses Coulee, a canyon in the Channeled Scablands of Washington state.

forces required to erode blocks of rock, and the notion that the floods were just large enough to erode their beds. It will also be challenging to confirm that similarly modest floods formed other Channeled Scabland canyons, because not all canyons contain features that record the progress of canyon incision in the same way as the well-preserved terraces in Moses Coulee. However, observations of erosion by smaller, modern floods<sup>9</sup> support the principles behind the authors' approach.

Larsen and Lamb's results raise the possibility that the largest known floods in the Solar System were smaller than previously estimated. Numerous floods crossed the surface of Mars during the past few billion years, carving enormous canyons that dwarf the Channeled Scablands. The source of the flood water remains a mystery, but each flood probably originated either when water erupted from an underground aquifer, or when a surface reservoir, perhaps created by melting ice, suddenly drained — a scenario similar to that of Lake Missoula. Brimful flow rates estimated from high-water marks in the biggest Martian canyons are tens of times greater than the largest estimates for the Channeled Scablands<sup>10</sup>. The immensity of these floods is even more shocking given the cold, dry conditions that have characterized the surface of Mars for at least the past 2 billion years.

Larsen and Lamb do not attempt to model the Martian floods, but their results support previous suggestions<sup>11,12</sup> that the canyons on Mars could have been carved by a succession of smaller floods. Such a scenario could help to resolve the discrepancy between flow rates

estimated from canyon topography and geological constraints on water supply rates<sup>11</sup>. A succession of floods would have required repeated replenishment of the water source, which has implications for Mars's ancient

#### AGEING

## Measuring our narrow strip of life

**In line with previous research, a demographic analysis corroborates the presence of a limit to human lifespan, indicating that increases in life expectancy are likely to slow down or stop over the coming years. [SEE LETTER P.257](#)**

S. JAY OLSHANSKY

**T**he British author Annie Besant once wrote<sup>1</sup>: “out of the darkness of the womb, into the darkness of the grave, man passes across his narrow strip of life.” The ration of time allocated to humans is of profound personal and scientific interest. On page 257, Dong *et al.*<sup>2</sup> turn to the demographic literature to analyse whether there is a limit to human lifespan — and find evidence to suggest that there is.

Before discussing the study at hand, we should define some relevant terms. Lifespan describes how long an individual lives. Life expectancy is a population-based estimate

of expected duration of life for individuals at any age, based on a statistical ‘life table’. And maximum lifespan is the age reached by the longest-lived member of a species.

**J. Taylor Perron** is in the Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.  
**Jeremy G. Venditti** is in the Department of Geography, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.  
e-mails: [perron@mit.edu](mailto:perron@mit.edu); [jeremy\\_venditti@sfu.ca](mailto:jeremy_venditti@sfu.ca)

1. Larsen, I. J. & Lamb, M. P. *Nature* **538**, 229–232 (2016).
2. Bretz, J. H. *J. Geol.* **31**, 617–649 (1923).
3. Pardee, J. T. *Geol. Soc. Am. Bull.* **53**, 1569–1600 (1942).
4. Baker, V. R. *Geol. Soc. Am. Spec. Pap.* **144**, 1–73 (1973).
5. Waitt, R. B. *J. Geol.* **88**, 653–679 (1980).
6. O'Connor, J. E. & Baker, V. R. *Geol. Soc. Am. Bull.* **104**, 267–279 (1992).
7. Wohl, E. E. in *Large Rivers: Geomorphology and Management* (ed. Gupta, A.) 29–44 (Wiley, 2007).
8. Clarke, G. K. C., Leverington, D. W., Teller, J. T., Dyke, A. S. & Marshall, S. J. *Quat. Sci. Rev.* **24**, 1533–1541 (2005).
9. Lamb, M. P. & Fongstad, M. A. *Nature Geosci.* **3**, 477–481 (2010).
10. Baker, V. R. *Nature* **412**, 228–236 (2001).
11. Manga, M. *Geophys. Res. Lett.* **31**, L02301 (2004).
12. Andrews-Hanna, J. C. & Phillips, R. J. *J. Geophys. Res.* **112**, E08001 (2007).

of expected duration of life for individuals at any age, based on a statistical ‘life table’. And maximum lifespan is the age reached by the longest-lived member of a species.

Human life expectancy has risen fairly steadily and rapidly over the past 150 years<sup>3</sup> in most countries. In 1990, colleagues and I predicted that this increase would slow over time<sup>4</sup>, and this has proved to be the case<sup>5</sup>. Maximum lifespan also seems to have risen steadily<sup>6</sup>, but this too might have reached an upper asymptote — no one is known to have lived longer than Jeanne Calment, who died in 1997 at the age of 122. Thus, the debate about life's limits is ongoing.

Some scientists speculate that fixed limits to





**Figure 1 | A limit to lifespan?** At 116, Emma Morano is the oldest known person alive today. Dong *et al.*<sup>2</sup> provide evidence that we are approaching the natural limit to human lifespan.

life are unlikely to exist, because they cannot be observed using the tools of mathematical demography<sup>7</sup>. Others suggest that unknown technological advances in the future will continue to drive down death rates<sup>8</sup>, leading to accelerated gains in life expectancy and maximum lifespan. And yet others argue that there is a limit to lifespan<sup>9</sup>.

Dong and colleagues used demographic data to investigate whether there is a limit to human lifespan and, by implication, life expectancy. They first hypothesized that, if a biological limit does not exist (or is currently unobservable), the age group experiencing the greatest increase in survival should shift to ever-older groups over time. This hypothesis makes perfect sense, and the authors discovered that, in most countries that have reliable data, the greatest improvement in survival in the oldest age groups peaked in about 1980 and has not shifted since.

Next, the researchers investigated whether increases in maximum lifespan had been observed in recent decades. They discovered that, since the death of Calment, maximum lifespan for humans has regressed. This occurred in spite of the increasing size of ageing populations worldwide, which, in itself, should have led to an increase in maximum lifespan. Dong *et al.* conclude that these two observations represent compelling evidence that human lifespan has a 'natural limit' (Fig. 1).

Scientists who study ageing know that there is considerable variation in the duration of life across species<sup>10</sup>, but within species there are fixed attributes associated with life history — and longevity determination is one of them. Under protected living conditions in which predation is largely removed, mice tend to live

about 1,000 days<sup>10</sup>, dogs about 5,000 days<sup>10</sup> and humans about 29,000 days<sup>11</sup>. Clearly, there are biological reasons for each species' average lifespan, so why would anyone think that people could live for much longer than we do now?

The answer lies in the historical context within which human longevity has changed. The 30-year rise in life expectancy at birth seen during the past century has nothing to do with a modified rate of ageing<sup>12</sup>. Instead, it reflects improvements in public health that have drastically reduced early-age mortality, allowing most people in developed nations to reach old age for the first time in history. Death now clusters in people between the ages of 65 and 95 (ref. 11). But, without further biomedical breakthroughs, life expectancy cannot continue to rise by much, and so future longevity gains will diminish. The crucial question is how much more survival time can be gained through medical technology. With fixed life-history traits, it would seem that we are running up against a formidable barrier.

As the authors rightly point out, the idea of a 'natural limit' to life does not imply that such a limit is a direct by-product of some genetically driven program that causes both ageing and death. Fixed genetic programs that directly cause ageing and death cannot exist as a direct product of evolution, because the end result would be death at an age beyond which almost every member of a species would ordinarily live. A genetic time bomb designed to kill us at older ages is equivalent to automobile manufacturers building in an explosive device that is set off only when a car reaches one million miles. Because most cars are never driven that far, such a device would be useless.

How is it possible to have a biological limit to

life, yet no genetic program that runs it? There are biological clocks that measure time from conception and birth, but these metronomes are there to transform a fertilized egg into an adult capable of reproducing. These fixed genetic programs for growth, development, maturation and reproduction (collectively known as a life-history strategy) are products of more than 3.7 billion years of evolution. Biological metronomes do not measure the time to age or die; instead, ageing is an inadvertent by-product of these clocks, which are designed to sustain life.

This distinction is important — it means that there is no fixed limit beyond which humans cannot live, but that there are, nevertheless, limits on the duration of life that are imposed by other genetically determined life-history traits. Think of constraints on running speed as an analogy. No genetic program specifically limits how fast humans can run, but biomechanical constraints on running speed are imposed by a fixed body design that evolved for other purposes. The absence of ageing and death programs opens the door to non-genetic interventions that extend health and length of life, just as new training methods enable us to run incrementally faster. This is precisely why modifying behavioural risk factors such as diet and exercise does extend the period of healthy life, but yields diminishing gains in life expectancy.

Dong and colleagues remind us that humanity is approaching a natural limit to life. This limit is now apparent in national vital statistics. Humanity is working hard to manufacture more survival time, with some degree of success, but we should acknowledge that a genetically determined fixed life-history strategy for our species stands in the way of radical life extension. ■

**S. Jay Olshansky** is in the UIC School of Public Health, University of Illinois at Chicago, Chicago, Illinois 60612, USA, and at Lapetus Solutions, Wilmington, North Carolina. e-mail: sjayo@uic.edu

1. Besant, A. in *The Origins of Theosophy: Annie Besant — The Atheist Years* (Routledge Revivals, 2015).
2. Dong, X., Milholland, B. & Vijg, J. *Nature* **538**, 257–259 (2016).
3. Riley, J. C. *Rising Life Expectancy: A Global History* (Cambridge Univ. Press, 2001).
4. Olshansky, S. J., Carnes, B. A. & Cassel, C. *Science* **250**, 634–640 (1990).
5. Crimmins, E. *Gerontologist* **55**, 901–911 (2015).
6. Wilmoth, J. R., Deegan, L. J., Lundström, H. & Horiuchi, S. *Science* **289**, 2366–2368 (2000).
7. Vaupel, J. W. *Nature* **464**, 536–542 (2010).
8. Christensen, K., Doblhammer, G., Rau, R. & Vaupel, J. W. *Lancet* **374**, 1196–1208 (2009).
9. Carnes, B. A., Olshansky, S. J. & Hayflick, L. *J. Gerontol. A* **68**, 136–142 (2013).
10. Austad, S. N. *J. Comp. Pathol.* **142** (Suppl. 1), S10–S21 (2010).
11. Human Mortality Database. <http://www.mortality.org/>
12. Olshansky, S. J. in *Ageing: The Longevity Dividend* (eds Olshansky, S. J., Martin, G. M. & Kirkland, J. L.) 221–237 (Cold Spring Harb. Lab. Press, 2016).

This article was published online on 5 October 2016.



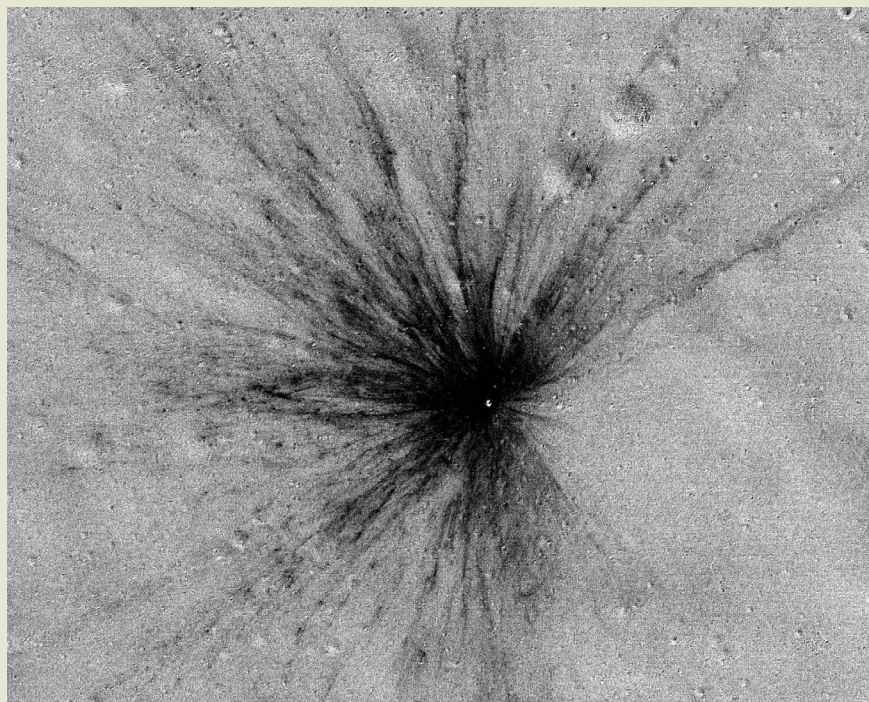
## PLANETARY SCIENCE

# Moon churn

The Moon's surface is being mapped by NASA's Lunar Reconnaissance Orbiter spacecraft, to aid planning for future missions. On page 215, Speyerer *et al.* report how images taken by the orbiter's camera have been used to quantify the current rate at which lunar craters form as a result of surface impacts by comets, asteroids and associated fragments (E. J. Speyerer *et al. Nature* **538**, 215–218; 2016).

The authors compared pairs of images of the Moon's surface taken at different times, and discovered that 222 craters had formed in the periods between the images being taken. They therefore estimate that about 180 craters of at least 10 metres in diameter form annually across the entire Moon. This is 33% more than would have been expected from a commonly used model of impact frequency.

By calculating the ratios of surface reflectance between pairs of images, Speyerer *et al.* uncovered distinct zones of subtly modified reflectance around the newly formed craters. The zones extend many crater widths out from the centre and are not visible to the naked eye (pictured are the ratios for a 12-m crater; dark regions reveal a zone that splays out up to 1,800 m from



the centre). The authors propose that these zones are caused by impact-induced jets of melted and vaporized material formed early in the crater-formation process.

The researchers also detected thousands of subtle surface disturbances — changes in local reflectance that lack a resolvable crater rim. They interpret many of these as the scars

of secondary impacts that churned up the upper few centimetres of the surface without forming a resolvable crater. Speyerer and colleagues therefore propose that the upper 2 centimetres of loose surface material on the Moon will be reworked in about 81,000 years, 100 times faster than previously predicted.

Andrew Mitchinson

## In retrospect

# Fifty years of C<sub>4</sub> photosynthesis

**Half a century after the discovery of a plant photosynthetic pathway termed C<sub>4</sub>, researchers are working to engineer this efficient pathway into crops such as rice to maintain food security.**

JULIAN M. HIBBERD & ROBERT T. FURBANK

Fifty years ago, Hatch and Slack<sup>1</sup> published an analysis of photosynthesis that gave birth to a new field. Their work not only stimulated intense biochemical research to define the mechanisms of a new photosynthetic pathway, but also fed into many other disciplines. Ecologists found that the pathway could explain species distributions. Geologists gained greater insight into changes in the isotope composition of sediments and fossils. And evolutionary biologists started to investigate the highly complex pathway, which is found in many plant lineages and is

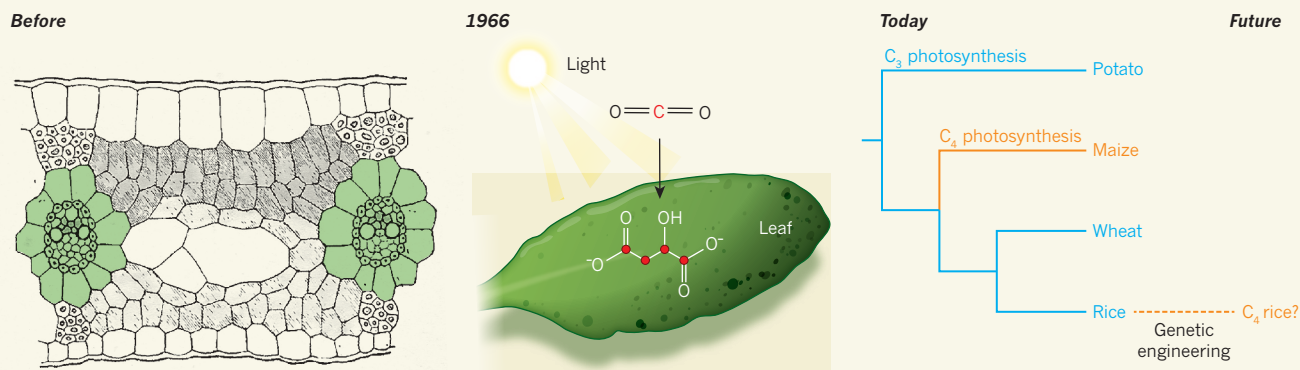
now considered one of the most remarkable examples of convergent evolution — a process in which the same feature evolves independently in different unrelated species.

Fifteen years before Hatch and Slack's work, Calvin and co-workers had identified the first photosynthetic pathway by which inorganic atmospheric CO<sub>2</sub> is incorporated (fixed) into organic carbon-containing molecules<sup>2</sup>. The initial step in the pathway produces a molecule that contains three carbon atoms, and it was widely thought that all land-dwelling plants used this 'C<sub>3</sub>' photosynthesis. However, this assumption was disproved by Hatch and Slack's carefully executed experiments. They used the carbon-14

isotope to create <sup>14</sup>CO<sub>2</sub> and then tracked how the <sup>14</sup>C was incorporated into molecules in sugarcane plants. Remarkably, they found that the first step of carbon fixation was actually into a four-carbon molecule<sup>1</sup>. This alternative pathway became known as C<sub>4</sub> photosynthesis. At the time, the significance of Hatch and Slack's finding was that two photosynthetic pathways were now known to operate in plants.

The study by Hatch and Slack explained some puzzling reports. Laboratories as far apart as Hawaii and Russia had observed unexpected carbon incorporation patterns when <sup>14</sup>CO<sub>2</sub> was supplied to sugarcane and maize (corn) leaves<sup>3,4</sup>. However, Calvin and others questioned the validity of those reports, and the findings were not accepted by the field. The main objection was that <sup>14</sup>CO<sub>2</sub> had often been introduced to leaves in the dark, when photosynthesis is not active, which risks creating artefacts of non-photosynthetic metabolism.

Hatch and Slack's key advance was providing a pulse of <sup>14</sup>CO<sub>2</sub> to leaves in light, followed by introduction of CO<sub>2</sub> that did not contain <sup>14</sup>C. Such 'pulse chase' experiments can track a <sup>14</sup>C wave as it transits through molecules in a pathway. The approach showed that the carbon was first incorporated into malate (Fig. 1), a molecule containing four carbons,



**Figure 1 | C<sub>4</sub> photosynthesis.** More than 100 years ago, circular structures known as Kranz anatomy (highlighted in green in the cross-sectional image<sup>16</sup> on the left) were observed in leaves, but their role was unknown. In 1966, Hatch and Slack used the <sup>14</sup>C isotope to track the fate of the carbon (red) of carbon dioxide as it was incorporated into intracellular organic carbon molecules during photosynthesis in leaves illuminated by light<sup>1</sup>. They found that the <sup>14</sup>C was incorporated into a malate molecule

containing four carbon atoms. This four-carbon molecule pointed to a new photosynthesis pathway, which was termed C<sub>4</sub> photosynthesis. Although C<sub>4</sub> photosynthesis has evolved independently many times across the plant phylogeny, including in maize (corn), many key crops such as potatoes, wheat or rice use a less efficient photosynthesis pathway called C<sub>3</sub>. A future goal is to try to engineer this efficient C<sub>4</sub> photosynthesis pathway into other plants such as rice (or potentially other C<sub>3</sub> crop plants).

and then transferred to the three-carbon molecule 3-phosphoglycerate (also produced by the enzyme RuBisCO in the C<sub>3</sub> pathway). This demonstration of carbon flux is still the most definitive evidence for a C<sub>4</sub> photosynthetic pathway.

Even before Hatch and Slack's discovery, there were clues that the physiology of some plants was different<sup>5</sup>. Tropical grasses often grow substantially faster, have higher photosynthetic rates and use water more efficiently than other plants. Furthermore, peculiar circular structures had been spotted more than 100 years ago in some leaves<sup>5</sup>. These structures, known as 'Kranz anatomy' (*Kranz* is the German word for wreath), are concentric circles of mesophyll cells that surround bundle sheath cells around the veins. The role of Kranz anatomy became apparent only in the context of C<sub>4</sub> photosynthesis.

By the 1980s, the fundamentals of the specialized biochemistry and modified anatomy of a C<sub>4</sub> leaf were known. All major enzymes required for the C<sub>4</sub> cycle had been identified, and the requirement for the C<sub>4</sub> pathway to be compartmentalized into two different cell types had been linked to Kranz anatomy. The C<sub>4</sub> leaf became a model for understanding specific cells of plants<sup>6</sup>, and researchers proposed that Kranz anatomy developed as a result of a diffusible molecule emanating from veins<sup>7</sup>.

In the decades following Hatch and Slack's advance, C<sub>4</sub> photosynthesis inspired other studies in diverse fields. In agriculture, it explained the high rates of photosynthesis and low water loss of some crops. The key factor is the PEPC enzyme, which acts at the start of the C<sub>4</sub> photosynthesis pathway. PEPC has a higher binding affinity for carbon than does the RuBisCO enzyme that acts at a similarly early step in C<sub>3</sub> photosynthesis. This means that the stomatal pores that allow atmospheric CO<sub>2</sub> to enter the leaf don't need to open as wide in a C<sub>4</sub> plant<sup>8</sup>, so water loss through the

stomata is reduced.

PEPC incorporates <sup>13</sup>C more readily than RuBisCO does<sup>9</sup>, and the resulting differences in leaf carbon-isotope signatures allows species classification as C<sub>3</sub> or C<sub>4</sub> plants in living tissue or fossils. This approach soon piqued the interest of ecologists and evolutionary biologists.

Ecologists realized that there were clear geographical distribution gradients of C<sub>3</sub> and C<sub>4</sub> plants, with C<sub>4</sub> plants dominating open habitats in the tropics and subtropics and gradually becoming less common farther from the equator<sup>10</sup>. The distribution was linked to the selective pressures that drive C<sub>4</sub> evolution, with C<sub>4</sub> plants commonly associated with dry conditions. Fossil studies provided insight into ancient environments. The <sup>13</sup>C content of C<sub>4</sub> leaves led to the discovery<sup>11</sup> that C<sub>4</sub> grasses rapidly expanded and came to dominate prairies and savannahs around 10 million to 6 million years ago: animals grazing on savannah grasses produce fossils that have a C<sub>4</sub>-type carbon-isotope signature; those browsing on C<sub>3</sub> trees produce fossils that have a C<sub>3</sub> isotope signature.

The discovery that C<sub>4</sub> plants arose across evolutionarily distantly related species implies repeated evolution of this complex photosynthetic pathway from the ancestral C<sub>3</sub> system. Despite the high complexity of the C<sub>4</sub> system, it has evolved independently many times, and the number of unrelated groups of C<sub>4</sub> plants has grown steadily, to around 61 lineages<sup>12</sup>. Several factors might have facilitated the repeated evolution of C<sub>4</sub> photosynthesis. For example, proteins of the C<sub>4</sub> pathway seem to be present in the ancestral C<sub>3</sub> state<sup>13</sup>. That means that, in principle, C<sub>4</sub> photosynthesis harnesses components already found in C<sub>3</sub> plants. Indeed, part of the C<sub>4</sub> pathway operates in some C<sub>3</sub> tissues<sup>14</sup>.

The repeated evolution of the C<sub>4</sub> system continues to intrigue. Although C<sub>4</sub> species are widely distributed within flowering plants, there are very few C<sub>4</sub> trees, a phenomenon that

has not yet been fully explained. For nearly 40 years, the C<sub>4</sub> pathway was thought to be dependent on pathway compartmentalization between mesophyll and bundle sheath cell types in the Kranz anatomy, but then several single-celled C<sub>4</sub> plants were discovered that broke that rule<sup>15</sup>.

This year, Hatch and Slack attended a fiftieth anniversary conference of their discovery in Canberra, Australia. They attended every session and were clearly excited, entertained and at times probably bemused by the variety of work now going on as a consequence of their work. Photosynthesis is approximately 50% more efficient in C<sub>4</sub> plants than in C<sub>3</sub> species<sup>8</sup>, leading to substantial yield benefits. Rice, the staple crop for more than half of the world's population, uses the C<sub>3</sub> pathway. As a consequence, substantial international efforts are now under way to understand the highly complex C<sub>4</sub> system sufficiently to allow it to be engineered into C<sub>3</sub> crops to increase their yield. ■

**Julian M. Hibberd** is in the Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK. **Robert T. Furbank** is in the Research School of Biology, Australian National University, Canberra, ACT 2601, Australia.

e-mails: [jmh65@cam.ac.uk](mailto:jmh65@cam.ac.uk); [robert.furbank@anu.edu.au](mailto:robert.furbank@anu.edu.au)

1. Hatch, M. D. & Slack, C. R. *Biochem. J.* **101**, 103–111 (1966).
2. Calvin, M. et al. *Symp. Soc. Exp. Biol.* **5**, 284–305 (1951).
3. Kortschak, H. P., Hartt, C. E. & Burr, G. O. *Plant Physiol.* **40**, 209–213 (1965).
4. Karpilov, Iu. S. *Tr. Kazan. Sel'shokhoz. Inst.* **41**, 15–24 (1960).
5. Furbank, R. T. *J. Exp. Bot.* **67**, 4057–4066 (2016).
6. Hibberd, J. M. & Covshoff, S. *Annu. Rev. Plant Biol.* **61**, 181–207 (2010).
7. Langdale, J. A. & Nelson, T. *Trends Genet.* **7**, 191–196 (1991).
8. Long, S. P. In: *C<sub>4</sub> Plant Biology* (eds Sage, R. F. &



Monson, R. K.) 215–249 (Academic Press, 1999).  
 9. O’Leary, M. H. *Phytochemistry* **20**, 553–567 (1981).  
 10. Teeri, J. A. & Stowe, L. G. *Oecologia* **23**, 1–12 (1976).  
 11. Cerling, T. E., Wang, Y. & Quade, J. *Nature* **361**, 344–345 (1993).

12. Sage R. F. *J. Exp. Bot.* **67**, 4039–4056 (2016).  
 13. Aubry, S., Brown, N. J. Hibberd, J. M. *J. Exp. Bot.* **62**, 3049–3059 (2011).  
 14. Hibberd, J. M. & Quick, W. P. *Nature* **415**, 451–454 (2002).

15. Voznesenskaya, E. V., Franceschi, V. R., Kiirats, O., Freitag, H. & Edwards, G. E. *Nature* **414**, 543–546 (2001).  
 16. Haberlandt, G. *Physiologische Pflanzenanatomie* (Engelmann, 1904).

## POPULATION GENETICS

# A map of human wanderlust

**Genetic studies of individuals from geographically diverse human populations provide insights into the dispersal of modern humans across the globe and how geography shaped genomic variation. SEE ARTICLES P.201 & P.207 & LETTER P.238**

SERENA TUCCI & JOSHUA M. AKEY

A remarkable feature of modern humans is our wanderlust, which the poet Charles Baudelaire famously referred<sup>1</sup> to as “*l’horreur du domicile*”. From our evolutionary birthplace in Africa<sup>2</sup>, modern humans have migrated to nearly every habitable corner of Earth (Fig. 1), overcoming obstacles such as ice, deserts, oceans and mountains. The number, timing and routes of human dispersals out of Africa have implications for understanding our past and how that past influenced contemporary patterns of human genomic variation. Three studies on pages 207, 201 and 238 (Malaspinas *et al.*<sup>3</sup>, Mallick *et al.*<sup>4</sup> and Pagani *et al.*<sup>5</sup>) describe 787 new, high-quality genomes of individuals from geographically diverse populations, providing opportunities to refine and extend current models of historical human migration.

In the past decade, the maturation of whole-genome sequencing technology has enabled data to be generated on a scale that was previously difficult to imagine. Genome-scale studies in humans, such as the 1000 Genomes Project<sup>6</sup>, which was completed last year, have contributed to a catalogue of genetic variation and genomic regions that confer the ability to adapt to diverse environments. Nonetheless, existing genetic data are often constrained by several factors, including limited breadth of population sampling and low-coverage data (in which each region of the genome is sequenced only a few times, leading to high error rates and missed variants). To address this issue, the current studies collect high-coverage sequence data for individuals from more than 270 populations across the globe. By studying the genetic diversity within and between these populations, the groups can tackle many questions about our past.

Cataloguing genetic data from indigenous populations, which are often difficult to access and are rapidly disappearing, is an important achievement. Mallick *et al.* and Pagani *et al.*

made great efforts to comprehensively sample regions that are typically under-studied; these include African populations, which have considerable genetic, linguistic and cultural diversity. Similarly, Malaspinas *et al.* describe the first extensive survey of human genetic diversity in Australia — a poorly studied region that, together with New Guinea, contains some of the earliest archaeological and fossil evidence of modern humans outside Africa.

The high-resolution portrait of human genetic diversity afforded by these studies allows new inferences to be made about our migration out of Africa. There are currently two conflicting models for such human dispersal. The first hypothesizes a single event that occurred around 40,000–80,000 years ago. Under this scenario, all present-day non-Africans trace their ancestry to a single population. By contrast, the multiple-dispersal model<sup>7</sup> posits that an initial migration out of Africa occurred as early as 120,000–130,000 years ago<sup>8</sup>, culminating in the peopling of southeast Asia and Australasia, possibly via a southern migration

route along the coastline of the Arabian peninsula and the Indian subcontinent. This early dispersal was followed by a second migration from Africa, through the Levant, which resulted in the peopling of mainland Eurasia.

Superficially, the current studies seem to come to different conclusions about out-of-Africa dispersals. Pagani *et al.* found that about 2% of genomes from individuals of Papua New Guinean ancestry indicate that their ancestors separated from Africans earlier than did other Eurasians. This observation is consistent with a multiple-dispersal model in which an early expansion of modern humans from Africa led to the peopling of Australasia around 120,000 years ago. This early out-of-Africa migration would have been followed by subsequent dispersals, and would have contributed only a small amount of ancestry to present-day Papuan individuals. Cranial morphology and other genetic data also support the idea of an early expansion<sup>9</sup>.

Malaspinas *et al.* and Mallick *et al.* consider a different sequence of events, in which all contemporary non-Africans branched off from a single ancestral population. Malaspinas and colleagues provide evidence that, on leaving Africa, modern humans immediately separated, leading to two waves of dispersal. As previously proposed<sup>10</sup>, one wave led to the peopling of Australasia, whereas the other contributed to the ancestry of present-day mainland Eurasians. Mallick and co-workers propose that this early separation instead occurred between west and east Eurasians, meaning that present-day people in Australia and Papua New Guinea might be descended from the same wave as east Asians.



**Figure 1 | A cave painting depicting human migration.** Three genetic studies<sup>3–5</sup> of individuals from geographically diverse populations provide clues about human history, including when and how many times humans moved out of Africa and throughout the world.

## ANIMAL BEHAVIOUR

However, neither Mallick *et al.* nor Malaspina *et al.* exclude the possibility of multiple out-of-Africa dispersals. Indeed, their models are consistent with earlier dispersals, as long as these early voyagers made little or no contribution to the gene pool of contemporary non-African populations (which is essentially what Pagani *et al.* find). Studies of ancient DNA clearly show that large-scale population turnovers have happened throughout human history: populations that once lived in Eurasia, for example, vanished without a trace, except for their bones<sup>11,12</sup>. Thus, although some differences between the proposed models are yet to be reconciled, they are not as disparate as they might seem to be.

The three studies also provide resources to better define models of genetic mixing between modern humans and their archaic hominin relatives, such as Neanderthals and Denisovans. Malaspina and colleagues propose that the genomes of present-day Aboriginal Australians might harbour traces of an ancient liaison with an unknown hominin group. Although evidence for gene flow from an unknown hominin group is tentative, it highlights the potentially surprising things that can be learnt from a comprehensive sampling of human genomic variation.

These studies fill in some missing pieces in the puzzle of human history, but many fascinating questions remain. The continued sampling of human genomic diversity and the development of increasingly sophisticated statistical tools promise to reveal more secrets about our past. Nonetheless, it is crucial to recognize the limits of genetics. As previously pioneered<sup>13</sup>, the integration of data across traditionally distinct disciplines, such as linguistics, archaeology, anthropology and genetics, will be necessary to fully retrace the steps taken by early humans as they explored and colonized the world. ■

**Serena Tucci and Joshua M. Akey** are in the Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.  
e-mail: akeyj@uw.edu

- Baudelaire, C. *Journaux intimes* (Crès, 1920).
- Stringer, C. B. & Andrews, P. *Science* **239**, 1263–1268 (1988).
- Malaspina, A.-S. *et al. Nature* **538**, 207–214 (2016).
- Mallick, S. *et al. Nature* **538**, 201–206 (2016).
- Pagani, L. *et al. Nature* **538**, 238–242 (2016).
- 1000 Genomes Project Consortium. *Nature* **526**, 68–74 (2015).
- Lahr, M. M. & Foley, R. *Evol. Anthropol.* **3**, 48–60 (1994).
- Armitage, S. J. *et al. Science* **331**, 453–456 (2011).
- Reyes-Centeno, H. *et al. Proc. Natl Acad. Sci. USA* **111**, 7248–7253 (2014).
- Rasmussen, M. *et al. Science* **334**, 94–98 (2011).
- Fu, Q. *et al. Nature* **524**, 216–219 (2015).
- Fu, Q. *et al. Nature* **514**, 445–449 (2014).
- Cavalli-Sforza, L. L. *The History and Geography of Human Genes* (Princeton Univ. Press, 1994).

This article was published online on 21 September 2016.

# Lethal violence deep in the human lineage

Researchers estimate that the incidence of human lethal violence at the time of the origin of our species was about six times higher than for the average mammal, but about as violent as expected, given our great-ape ancestry. [SEE LETTER P.233](#)

MARK PAGEL

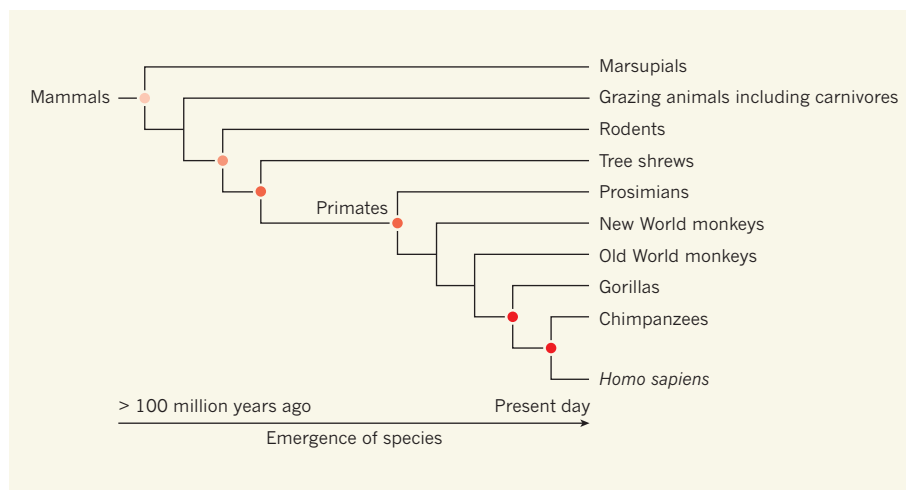
Are humans naturally violent, as the seventeenth-century philosopher Thomas Hobbes thought<sup>1</sup>, with the prevailing condition of humans being one of “continual feare, and danger of violent death”, or as Jean-Jacques Rousseau imagined<sup>2</sup> a century later, neither good nor bad but moulded by their environments? Social scientists have long confronted this question by estimating rates of human violence after controlling for factors such as age, sex, race and income in large cohorts of individuals drawn from a variety of circumstances. On page 233, Gómez *et al.*<sup>3</sup> adopt a different approach: they use comparative methods from evolutionary biology<sup>4</sup> to reconstruct probable ancestral rates of lethal violence at the time of the origin of our species roughly 160,000 to 200,000 years ago.

One of Charles Darwin’s great insights was that all living things evolve by a process of descent with modification, such that species give rise to daughter species that inherit many of their ancestors’ traits. Comparative biologists can use the family trees (phylogenies)

that arise from this process to infer the history of biological evolution, to date past events, and to reconstruct probable ancestral features of species that lived hundreds of thousands to hundreds of millions of years ago.

Gómez and colleagues applied comparative statistical techniques to a phylogeny of mammals, which includes primates — the grouping of mammals that comprises monkeys, great apes and the lineage that leads to modern humans. The authors compiled information on more than 4 million deaths from 1,024 mammalian species drawn from 137 mammalian families (80% of the total number of mammalian families), including mice, horses, bats, rabbits and monkeys. Information for humans came from 600 studies, and was derived from palaeolithic samples (defined by the authors as 50,000 to 12,000 years ago), New World and Old World Mesolithic (12,000 to 10,200 years ago) and Neolithic (10,200 to 5,000 years ago) sites, Bronze Age (5,300 to 3,200 years ago) and Iron Age (3,200 to 1,300 years ago) samples, and anthropological sources from the past few centuries.

The authors then calculated the proportion of deaths attributable to violence from a



**Figure 1 | Rates of lethal violence in a mammalian family tree.** Gómez *et al.*<sup>3</sup> used data about causes of death in more than 1,000 mammalian species to determine for each species the proportion of deaths caused by members of the same species. These data were grouped into a phylogenetic family tree for mammals, of which a simplified representation is shown here. The rates of lethal violence against members of the same species are indicated by a colour scale that ranges from low rates in light red to higher rates in darker red.



member of the same species out of all deaths counted for each species. Including only acts of within-species violence is a key point. Lions and tigers routinely kill members of other species, but they are less often lethally violent towards each other. Within-species violence therefore gets at what Hobbes and Rousseau disagreed about, and is interesting because members of the same species typically have all the same weapons, making violence between them risky.

Using the values from contemporary species, the authors reconstructed the rate of lethal violence (caused by members of the same species) at the phylogenetic origin of mammals at about 0.30%, which is approximately 1 in 300 deaths. Rates of lethal violence then rose steadily over time throughout the mammalian phylogeny (Fig. 1) as the reconstructed ancestors drew closer to primates: the rate is about 1.1% for the ancestor of primates, rodents and hares; 2.3% for the common ancestor of primates and tree shrews; then drops slightly to 1.8% for the ancestor of the great apes. The increases in lethal violence coincide with species having increasing amounts of group living and territoriality. Group living places individuals routinely in close contact, and territoriality means that groups might potentially compete over resources. Gómez and colleagues reconstructed the incidence of human lethal violence at the origin of our species at 2%, about six times higher than the reconstructed mammalian value.

Gómez and colleagues' study risks being misunderstood, so it is necessary to be clear about its interpretation. Humans emerged from an evolutionary lineage with a long history of higher-than-average levels of lethal violence towards members of the same species. Even so, followers of Rousseau might step in to say that our species' figure of 2% tells us nothing about our innate tendencies; it might merely reflect a calculated or environmentally induced response to the environments in which early humans lived.

Perhaps, but this objection falters when we appreciate that species that live in particular kinds of environments over long periods of time tend to adapt to those environments genetically, and this makes some kinds of outcomes more likely than others: a wolf raised as a sheep will probably one day turn on its fellow sheep. For this reason, the authors' finding of a steady increase in violence throughout the 100 million or so years of the mammalian tree is important — there was plenty of time for our ancestors to acquire and bequeath us genetic adaptations towards lethal violence. Our nearest living relatives, the chimpanzees, with whom we share around 98% identity of our gene sequences<sup>5</sup>, form packs to hunt down and kill stray males from other chimpanzee tribes<sup>6</sup>, and their hunting parties bear resemblances to human hunter-gatherer warfare<sup>7</sup>. Even the usually peaceful bonobo *Pan paniscus* can



**Figure 2 | Primate violence.** Aggressive behaviour can occur even in the normally peaceful bonobo, *Pan paniscus*.

sometimes display violent behaviour (Fig. 2).

Some will object that it is difficult to derive reliable estimates of lethal violence. Anticipating this, the authors test for several biases, including variation in sample sizes and sampling effort, and uncertainty about the phylogeny of mammals itself, showing that none of these qualitatively alters their results. They also find that species we expect to be violent, such as the predatory carnivores, are violent, and species that we do not expect to be violent because they are mainly vegetarian, tend not to be. Finally, the authors show that rates of violence are heritable in the mammalian phylogeny, by demonstrating that closely related species tend to have similar levels of lethal violence.

Still, the Rousseau camp might have a corner to fight. The authors' estimates of rates of lethal violence in humans vary widely over time, in most cases too quickly to be attributable to genetic changes. Their palaeolithic samples have rates very close to the 2% predicted at the origin of our species, but then rates rise to as high as 15–30% (with high statistical uncertainty) in samples from between 3,000 and 500 years ago, before declining in contemporary populations (approximately 100 years ago to the present day). The rise tends to correlate with moving from an early pre-societal 'state of nature' to tribal groupings and then to organized political societies that have a warrior class.

Where does this leave us? Social scientists take note: the work by Gómez and colleagues opens up a new approach to uncovering the origins of human violence, giving good

grounds for believing that we are intrinsically more violent than the average mammal, and their findings fit well with anthropological accounts that describe hunter-gatherer societies as being engaged in 'constant battles'<sup>8,9</sup>. But societies can also modify our innate tendencies. Rates of homicide in modern societies<sup>10</sup> that have police forces, legal systems, prisons and strong cultural attitudes that reject violence are, at less than 1 in 10,000 deaths (or 0.01%), about 200 times lower than the authors' predictions for our state of nature. Hobbes has landed a serious blow on Rousseau, but not quite knocked him out. ■

**Mark Pagel** is in the School of Biological Sciences, University of Reading, Reading RG6 6UR, UK.  
e-mail: m.pagel@reading.ac.uk

1. Hobbes, T. *Leviathan* (Penguin, 1968).
2. Rousseau, J.-J. *The Social Contract* (Wordsworth, 1998).
3. Gómez, J. M., Verdú, M., González-Megías, A. & Méndez, M. *Nature* **538**, 233–237 (2016).
4. Harvey, P. H. & Pagel, M. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, 1991).
5. The Chimpanzee Sequencing and Analysis Consortium. *Nature* **437**, 69–87 (2005).
6. Wilson, M. L. et al. *Nature* **513**, 414–417 (2014).
7. Wrangham, R. W. & Glowacki, L. *Hum. Nature* **23**, 5–29 (2012).
8. Keeley, L. H. *War Before Civilization* (Oxford Univ. Press, 1996).
9. LeBlanc, S. A. & Register, K. E. *Constant Battles* (St Martin's Press, 2003).
10. United Nations Office on Drugs and Crime. *Global Study on Homicide* (UNDOC, 2011); available at [go.nature.com/2dacpri](http://go.nature.com/2dacpri)

This article was published online on 28 September 2016.

# In vitro and ex vivo strategies for intracellular delivery

Martin P. Stewart<sup>1,2\*</sup>, Armon Sharei<sup>1,2,3,4\*</sup>, Xiaoyun Ding<sup>1,2</sup>, Gaurav Sahay<sup>5</sup>, Robert Langer<sup>1,2</sup> & Klavs F. Jensen<sup>1</sup>

**Intracellular delivery of materials has become a critical component of genome-editing approaches, ex vivo cell-based therapies, and a diversity of fundamental research applications. Limitations of current technologies motivate development of next-generation systems that can deliver a broad variety of cargo to diverse cell types. Here we review in vitro and ex vivo intracellular delivery approaches with a focus on mechanisms, challenges and opportunities. In particular, we emphasize membrane-disruption-based delivery methods and the transformative role of nanotechnology, microfluidics and laboratory-on-chip technology in advancing the field.**

Despite its essential role in biological research and therapeutic applications, the efficient intracellular delivery of exogenous compounds and macromolecular cargo remains a long-standing challenge. The limitations of established delivery technologies have hampered progress in multiple areas as the potential of exciting new materials, insights into disease mechanism, and approaches to cell therapy are not fully realized owing to their delivery hurdles. This challenge can be viewed through the lens of two broad parameters: cell type and target material. Existing technologies are mainly focused on addressing a subset of combinations, specifically nucleic acid delivery (that is, transfection) to immortalized cell lines and certain primary cells. Some of the most exciting target cell types, such as stem cells and immune cells, are also the most difficult to address. Thus, methods of delivering almost any cargo molecule to any cell type are much needed.

Although carrier-mediated delivery systems offer promise for nucleic acid transfection *in vivo*<sup>1,2</sup>, membrane-disruption-based modalities are attractive candidates for universal delivery systems *in vitro* and *ex vivo*. In this review, we begin with motivations driving next-generation intracellular delivery strategies and suggest relevant requirements for future systems. Next, a broad overview of current delivery concepts covering salient strengths, challenges and opportunities is presented. Following that, our focus shifts to prevalent mechanisms of membrane disruption and recovery in the context of intracellular delivery. Finally, we highlight the potentially transformative role of nanotechnology, microfluidics and laboratory-on-chip approaches in shaping the field.

## Next-generation intracellular delivery

Next-generation intracellular delivery solutions are required in diverse scenarios ranging from cell-based therapy and gene editing to regenerative medicine and fundamental biology (Fig. 1). *Ex vivo* cell-based gene therapies have shown promise in clinical trials against human disease, with exciting examples including self-renewing haematopoietic stem cells and T cells for immunotherapy<sup>3–6</sup>. In haematopoietic stem cells, gene therapy to correct mutations in monogenic diseases such as severe combined immunodeficiency (SCID)-X1, Wiskott–Aldrich Syndrome, and  $\beta$ -thalassemia has been achieved<sup>4</sup>. For T cells, novel function against tumour targets can be instructed by induced expression of specific T cell receptors and chimaeric antigen receptors followed by adoptive cell transfer<sup>5</sup>. Non-essential proteins used by pathogenic processes can

be deleted by delivery of genome-editing nucleases, such as was recently demonstrated with ablation of the HIV-dependent CCR5 receptor necessary for infection in T cells<sup>7</sup> or the haematopoietic lineage<sup>8</sup>. Moreover, induced secretion of cytokines, or programmed drug resistance and safety switches, can be engineered into these cell types by *ex vivo* manipulation<sup>4,6</sup>. Recent breakthroughs in gene editing with programmable nucleases offer an unparalleled opportunity to reach many of these goals<sup>9,10</sup>. Further afield, in regenerative medicine the importance of not relying on potentially mutagenic viral vectors for induced pluripotent stem cell production has spawned reprogramming efforts by direct delivery of proteins<sup>11</sup>, messenger (m)RNA transfection<sup>12</sup>, and micro (mi)RNA delivery<sup>13</sup>, among other possibilities. A common theme in these clinically relevant examples is the need to perform efficient and safe intracellular delivery.

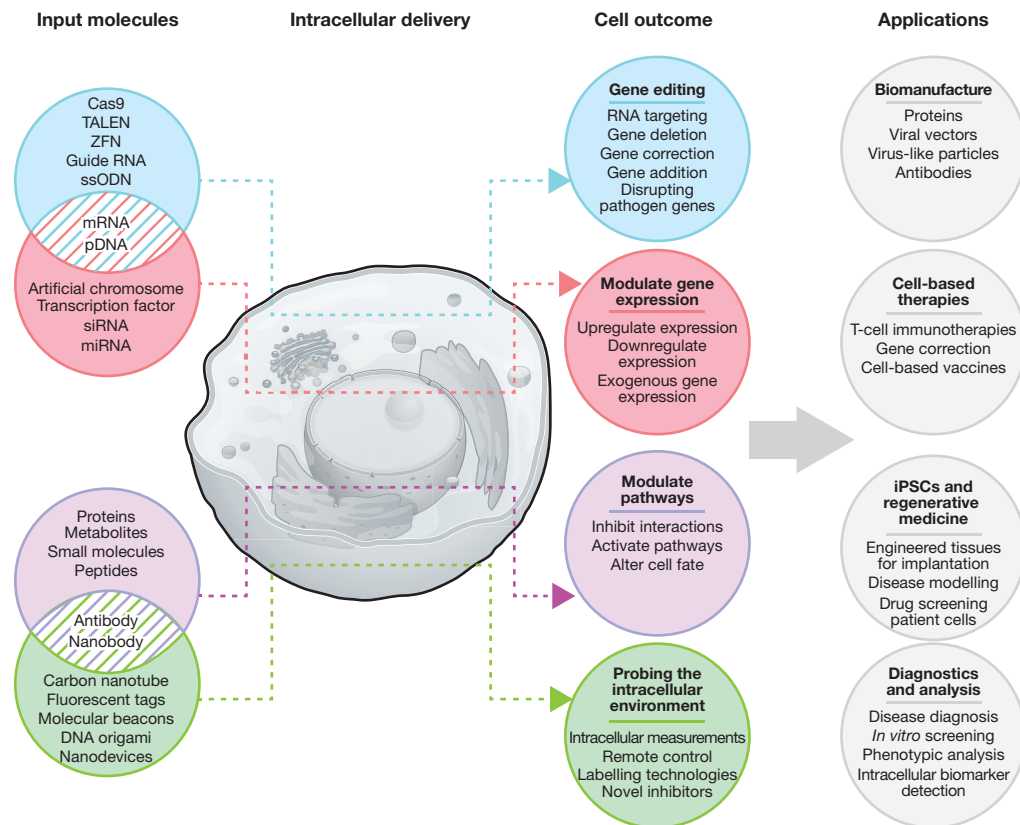
In basic research, nucleic acids, proteins, peptides, metabolites, membrane impermeable drugs, cryoprotectants, exogenous organelles, molecular probes, nanodevices and nanoparticles are all potential target materials for intracellular delivery. Despite their limitations, current delivery capabilities, largely centred around nucleic acid delivery to cell lines, have already yielded dramatic progress with plasmid (p)DNA and mRNA for gene expression and small interfering (si)RNA and miRNA for gene silencing. Meanwhile, systematic delivery of protein biologics into living cells, such as active inhibitory antibodies and stimulatory transcription factors, represent a powerful yet largely untapped tool for decoding and engineering cell function<sup>14</sup>. Measurement of intracellular chemical and physical properties with innovative devices, sensors and probes is another frontier<sup>15</sup>. Probes engineered from functional nanomaterials—including nanoplasmonic optical switches<sup>16</sup>, carbon nanotubes<sup>17</sup> and quantum dots<sup>18</sup>—have generated excitement in research communities for decades but ineffective intracellular delivery, a poor understanding of their interaction with biological environments, and toxicity issues have retarded their deployment in the cellular context. These delivery challenges are particularly acute in the case of important patient-derived cell types such as immune cells, stem cells and neurons<sup>19,20</sup>.

Taken together, the preceding examples (Fig. 1) provide compelling motivations and requirements for next-generation intracellular delivery systems. In Table 1 we propose a set of guidelines to be considered by inventors. Regarding cell-based therapies, for example, decades of

<sup>1</sup>Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. <sup>2</sup>The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. <sup>3</sup>Department of Microbiology and Immunobiology, Harvard Medical School, Boston, Massachusetts, USA. <sup>4</sup>The Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts, USA. <sup>5</sup>Department of Pharmaceutical Sciences, Collaborative Life Science Building, College of Pharmacy, Oregon State University, Portland, Oregon, USA.

\*These authors contributed equally to this work.





**Figure 1 | Intracellular delivery is a key step in investigating and engineering cells.** The schematic depicts examples of application areas and molecular tools that require intracellular delivery for their realization. The subsets are not mutually exclusive. For example, gene editing may

be employed in cell-based therapies, regenerative medicine, and genetic modulation. iPSC, induced pluripotent stem cell; TALEN, transcription-activator-like effector nuclease; ZFN, zinc finger nuclease; ssODN, single-stranded donor oligodeoxynucleotides.

clinical trials indicate that the risks of *ex vivo* culture include karyotype abnormalities, genotoxicity, and exhaustion of proliferative potential<sup>3</sup>. Hence rapid and safe delivery protocols that maximize efficiency and cell

viability while minimizing time in culture are crucial. For gene editing, a key issue is that transient 'hit and run' exposure of nucleases is often more favourable than indirect expression from DNA or mRNA, because

**Table 1 | Ideal features of next-generation intracellular delivery systems**

Feature	Justification
Minimal cell perturbation	<ul style="list-style-type: none"> <li>The exogenous vectors, materials or physical forces required to facilitate delivery can lead to off-target effects and toxicity</li> <li>Prolonged culture duration associated with delivery and verification can lead to unintended fate or phenotypic changes, loss of proliferative or homing potential, genotoxicity, and karyotyping abnormalities and accumulated mutations</li> <li>By minimizing the physical or biochemical manipulation necessary to achieve delivery, one can reduce undesirable side effects and maximize efficacy of the delivery process</li> </ul>
Scalability	<ul style="list-style-type: none"> <li>Effective delivery systems must be amenable to implementation at different scales of throughput</li> <li>Studying a rare cell subset may require only 100 cells per sample whereas an adoptive-transfer T-cell therapy can involve transfer of over <math>10^8</math> cells per patient</li> <li>Considering processes complying with current good manufacturing practice (cGMP) at an earlier stage may accelerate clinical translation</li> </ul>
Universal across cell types	<ul style="list-style-type: none"> <li>An ideal delivery system should be able to accommodate the diversity of physical and biological properties of potential target cells to ensure efficacy in the applications of interest; this could be accomplished through exploiting relatively universal mechanisms to facilitate delivery</li> </ul>
Material independent	<ul style="list-style-type: none"> <li>Delivery materials of interest may have diverse chemical and physical properties</li> <li>To facilitate robust delivery, an ideal delivery system should rely on a delivery mechanism that is independent of material properties (for example, if a delivery technology relies on electrophoresis to facilitate delivery it may not be compatible with uncharged materials)</li> </ul>
Compatible with intracellular targeting	<ul style="list-style-type: none"> <li>Distinct materials have different target sites within the cell (for example, siRNA and mRNA should facilitate the desired gene knockdown or expression effects in the cytosol, whereas DNA transcription requires nuclear localization)</li> <li>A preferred delivery system must be compatible with various intracellular targeting strategies; specifically, it should provide robust delivery of material to the cytosol and not interfere with targeting motifs (for example, a nuclear localization peptide sequence) on the material</li> </ul>
Dosage control	<ul style="list-style-type: none"> <li>The ability to control the maximum and minimum intracellular concentration may be key in some applications</li> <li>Strategies that rely on indirect delivery by expression from nucleic acids are subject to cell-to-cell variation and inherent signal amplification associated with transcription</li> </ul>
Cost	<ul style="list-style-type: none"> <li>Cost and complexity of production/operation can limit the utility of a delivery technology</li> <li>Effective delivery systems should use scalable, cost-effective designs that are amenable to clinical translation, compliance with cGMP standards, and large-scale manufacturing</li> </ul>

it yields more control over dosage concentration and efficacy, and reduces off-target effects<sup>21</sup>. For induced pluripotent stem cells, delivery challenges may impede the use of the ideal combination of proteins, nucleic acids, and small molecules to provide the optimal reprogramming outcome. Furthermore, low-cost transfection remains a barrier in many sensitive cell types at the level of basic research, and is even more problematic when considering scale-up for clinical protocols or industrial processes. Therefore, next-generation intracellular delivery strategies must strive to address the decades-old challenge of delivering diverse cargoes to the intracellular space of a wide range of cell types.

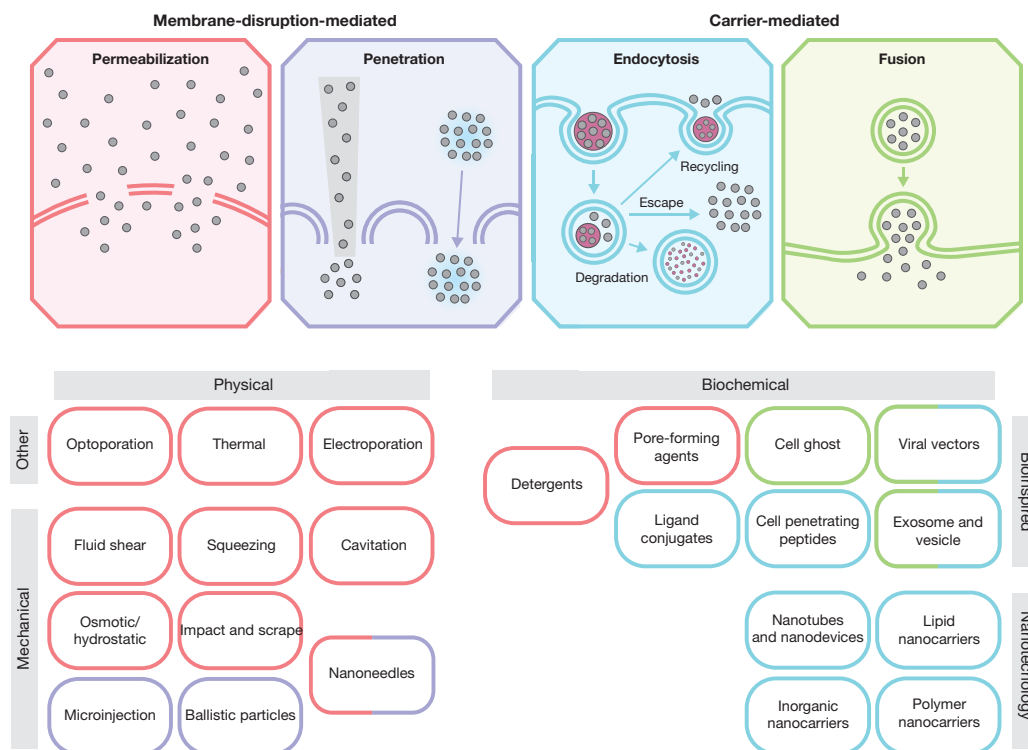
### Many roads to one destination

Intracellular delivery can be achieved by a range of carrier-based or membrane-disruption-based techniques (Fig. 2). Membrane-disruption modalities are primarily physical, involving the introduction of transient discontinuities in the plasma membrane via mechanical, electrical, thermal, optical or chemical means. These approaches can be thought of as permeabilization or direct penetration modalities. A cell becomes permeable to a substance when disruptions in the membrane are of sufficient size to allow passage through the membrane. Alternatively, direct penetration employs a solid conduit or vehicle to concurrently penetrate the membrane and introduce cargo. Carrier-based approaches comprise various biochemical assemblies, mostly of molecular to nanoscale dimensions. The purpose of carriers is threefold: (1) to package the cargo and protect it from degradation, (2) to gain access to the intended intracellular compartment, and (3) to release the payload with the appropriate spatiotemporal dynamics. Carriers can be bio-inspired, such as reconstituted viruses, vesicles, cell ghosts, and functional ligands and peptides. They may be based upon synthesis techniques from chemistry, materials science and nanotechnology, involving assembly of nanoparticles and macromolecular complexes from organic and inorganic origins. Most carriers enter through endocytosis but some may exhibit

fusogenic potential, endowing them with the ability to merge directly with the target membrane.

In the case of nucleic acid delivery (that is, transfection), vectors are defined as constructs that contain foreign DNA for the purpose of expression or replication. Major vector types are plasmids, cosmids, episomes/artificial chromosomes, and viral vectors, of which only viral vectors are capable of unassisted entry. Viral vectors exploit the viral infection pathway to enter cells but avoid the subsequent expression of viral genes that leads to replication and pathogenicity<sup>22</sup>. This is done by deleting coding regions of the viral genome and replacing them with the DNA to be delivered, which either integrates into host chromosomal DNA or exists as an episomal vector. At present, viral vectors are the most clinically advanced nucleic acid delivery agents owing to their high efficiency and specificity. They have been implemented in clinical trials for decades, being viewed as a promising approach for gene therapy<sup>22,23</sup>. In 2015, more than half of the gene therapy submissions to the Federal Drug Agency (FDA) relied on viral vector platforms based on lentivirus, retrovirus, adenovirus, or adeno-associated virus. For *ex vivo* applications, lentiviral transduction of the haematopoietic lineage is a prominent example<sup>3,4</sup>. However, challenges such as immune response, safety and complexity of preparation are concerns for viral vectors<sup>22,23</sup> and so viral systems have struggled to gain FDA approval. To address these issues, researchers are developing new vectors and serotypes<sup>4</sup>. Table 2 presents a comparison of strengths, challenges and opportunities for viral vectors, non-viral carriers and membrane-disruption-based delivery.

Motivated by the limitations of their viral counterparts, hundreds of non-viral vectors and synthetic carriers have been designed, using vast combinations of lipid, polymer, and inorganic nanomaterials, sometimes featuring functionalization with ligands, cell-penetrating peptides and other targeting or stabilizing agents (see reviews<sup>1,24–26</sup>). Most carriers are designed for nucleic acid transfection, but recent efforts seek to expand their ability to co-deliver proteins or other biomolecules<sup>27</sup>. Nearly all



**Figure 2 | Map of the relationship between intracellular delivery approaches, basic mechanism and conventional physical and biochemical categorizations.** Physical techniques produce membrane disruption either via permeabilization or direct penetration, while biochemical assemblies and viral vectors act as carriers to shuttle cargo through endocytosis. If a carrier has fusogenic potential, it may

also enter through membrane fusion. Some biochemical approaches, such as detergents and pore-forming proteins, work via membrane permeabilization. Schematics at the top show the four subcategories with molecular cargo (grey), membrane (double lines), and carrier material (purple).



**Table 2 | Strengths, challenges and opportunities for intracellular delivery approaches**

	Carriers		Membrane disruption
	Viral vectors	Non-viral	
Strengths	<ul style="list-style-type: none"> <li>• Amenable to both <i>in vivo</i> and <i>in vitro</i> translation</li> <li>• Experience gained from advanced status in clinical trials</li> <li>• High efficiency of intracellular delivery due to viral exploitation of infection pathway</li> </ul>	<ul style="list-style-type: none"> <li>• Amenable to both <i>in vivo</i> and <i>in vitro</i> translation</li> <li>• Packaging of delivery material can protect payload from premature degradation and potentially enable more efficient use of valuable materials</li> <li>• Capable of cell-specific and intracellular targeting</li> <li>• Passive, potentially high throughput</li> </ul>	<ul style="list-style-type: none"> <li>• Delivery of diverse materials</li> <li>• Capable of addressing many cell types</li> <li>• Optionally vector-free (that is, non-immunogenic)</li> <li>• Transient, defined exposure to membrane disruption</li> <li>• Rapid, almost instantaneous delivery</li> </ul>
Challenges	<ul style="list-style-type: none"> <li>• Can trigger adverse host immune response</li> <li>• Limited to nucleic acid delivery (transduction)</li> <li>• Limited genome size</li> <li>• Many are cell-cycle dependent</li> <li>• Preparation may be expensive, time consuming, require extensive experience, and demand special safety measures (for example, BL2)</li> <li>• For integrating viruses, risk of genotoxicity</li> <li>• Limited tropism may restrict target cell types</li> <li>• Manufacturing challenges for scaling up, for example, quality control for vector potency</li> </ul>	<ul style="list-style-type: none"> <li>• Inefficient and slow delivery, especially for carriers that enter via endocytosis (about 1% endosomal escape)</li> <li>• Carrier materials may perturb cell function or cause toxicity in unpredictable ways</li> <li>• <i>In vivo</i> targeting outside the liver has been difficult</li> <li>• Often restricted to delivering particular types of cargo (such as nucleic acids)</li> <li>• Unpacking kinetics may be unfavourable</li> <li>• Complex, laborious and expensive biochemistry or materials synthesis may be required for carrier preparation and manufacturing</li> </ul>	<ul style="list-style-type: none"> <li>• Loss of cytoplasmic content</li> <li>• Some modalities (such as thermal and electric) can lead to excessive damage to organic molecules, protein denaturation, and internal membrane breakdown</li> <li>• Some approaches (such as microinjection) currently not amenable to high throughput</li> <li>• Some methods can be restricted to adherent-only or suspension-only cells</li> <li>• Less amenable to <i>in vivo</i> translation</li> </ul>
Opportunities	<ul style="list-style-type: none"> <li>• Less-immunogenic vectors or evasion of immune response</li> <li>• Programmable tropism and specificity</li> <li>• Further development of hybrid viral serotypes</li> <li>• Potential of alternative viral species for new vector development</li> <li>• Improved production methods</li> </ul>	<ul style="list-style-type: none"> <li>• Novel carriers that can efficiently co-deliver diverse cargos (for example, proteins + nucleic acids)</li> <li>• Manipulation of target cell biology to regulate membrane trafficking and increase delivery efficiency</li> <li>• Stimuli-sensitive nanocarriers</li> <li>• Leveraging direct fusion to bypass endocytosis (for example, exosomes)</li> <li>• Implementation of biomimetic functionality inspired by viruses, bacterial machinery, and exosomes</li> </ul>	<ul style="list-style-type: none"> <li>• Microfabrication and nanotechnology enables fine control of physical phenomena and makes modalities that were previously intractable more feasible</li> <li>• A better understanding of cell recovery processes may allow mitigation of toxicity and functionality issues</li> <li>• Can be combined with carriers designed for subcellular targeting and controlled release</li> <li>• Engineered switchable valves in plasma membrane to dynamically control permeability</li> <li>• Versatile potential for <i>in vitro</i> and <i>ex vivo</i> applications</li> </ul>

See text for further details and references.

carriers are taken up via specific endocytic pathways based on their cell surface interactions and physicochemical properties<sup>28–30</sup>. To reach the intended intracellular target, the cargo must escape endosomal progression, which otherwise leads to degradation in lysosomes or regurgitation back to the cell surface<sup>31</sup>. For lipid nanocarriers, which are considered the most advanced non-viral vectors for nucleic acid delivery, quantitative studies reveal that approximately 1% of the nanocarriers escape from endosomes<sup>32</sup>. The exact mechanisms of escape remain elusive, however. Proposed explanations include endosome disruption, either by formation of transient lesions or vesicle lysis; active transport of dissociated products; or fusion of carriers or multi-vesicular bodies with the outer limiting membrane<sup>30–32</sup>. Apart from endosomal escape, another consideration is the kinetics of cargo release from the carrier, where delayed unpacking has been reported as a bottleneck to transfection efficiency<sup>33</sup>. Moreover, toxicity of carrier material and perturbation to membrane trafficking processes have been noted<sup>32,34</sup>. Manipulation of the host cell biology, using small molecules for example, represents an opportunity for boosting endosomal escape and delivery efficiency<sup>35</sup>. The design of stimuli-sensitive nanocarriers that respond to selective endosomal or intracellular conditions could also lead to improvements<sup>26</sup>.

Carriers with fusion capabilities circumvent endocytosis by releasing their cargo directly into the cytoplasm. These systems were first inspired by viruses that deploy specialized surface proteins to induce fusion with target membranes<sup>36,37</sup>. Fusogenic carriers are bound by a phospholipid bilayer that hosts the fusion machinery. Examples include cell ghosts, dead cells that have their cytoplasm replaced with cargo<sup>36,37</sup>, and virosomes, loaded vesicles reconstituted to present functional viral proteins<sup>38</sup>. More recently, cell-derived vesicles known as exosomes have been discovered to fuse with target cell membranes for the exchange of RNA and proteins

between immune cells<sup>39</sup>. Although the exact fusion mechanisms are yet to be elucidated, such bioinspired systems may represent a new generation of vehicles with which to overcome the poor efficiency and toxicity of synthetic carriers<sup>40</sup>.

An inherent limitation of carrier systems is the restricted combination of feasible cargo materials and cell types. Target cells may not exhibit the appropriate receptors, surface interactions, endocytic activity, or endosomal escape pathways. Furthermore, potential cargo materials often display enormous variability in their properties, such as charge, hydrophobicity, size, mechanical properties, composition, and functional groups. They may not efficiently complex with the carrier, tolerate packaging, unpack properly, or be amenable to delivery in sufficient quantities for a given application. For example, cationic lipids readily form complexes with anionic nucleic acids to transfect most immortalized cell lines, but many blood and immune cells remain recalcitrant<sup>20,41</sup>. On the other hand, cell-type-specific uptake can be a characteristic deliberately employed to achieve controlled targeting of a cell population<sup>42</sup>. Furthermore, owing to the mechanism of passive dispersion, carrier-based delivery is scalable and amenable to high throughput, with the ability to concentrate and protect limited amounts of cargo material for potent delivery. Related to these strengths, lipid nanoparticles and more minimalistic, compact conjugates are now in human clinical trials for *in vivo* delivery of therapeutic siRNAs<sup>1,2</sup>.

Unlike carriers, membrane-disruption-based approaches are less dependent on cargo properties, being able to deliver almost any sub-micrometre material dispersed in solution. The ability to rapidly switch membrane-perturbing effects on and off enables temporal control and rapid, almost instantaneous delivery. A further strength of membrane-disruption techniques *in vitro* and *ex vivo* is the broad range of cell types

and materials that can be addressed. Electroporation, for example, has a reputation for transfecting primary cell types that are otherwise recalcitrant to lipid nanoparticles and other non-viral transfection agents<sup>43</sup>. Membrane-disruption-based approaches may furthermore be combined with carriers to synergize the strengths of both, such as by delivering a nuclear-targeted DNA lipoplex to the cytoplasm<sup>44</sup>. Membrane-disruption-based delivery has also enabled several protein-delivery applications, featuring antibodies, transcription factors and genome-editing nucleases<sup>14,21,45,46</sup>. In primary human haematopoietic stem cells and T cells, for example, it was found that expression of Cas9 nuclease and guide RNA from plasmids was poorly tolerated, while direct delivery of Cas9–sgRNA complexes via electroporation improved efficiency, reduced off-target effects and normalized dosage control<sup>21,46</sup>. Such results highlight a trend towards direct delivery of macromolecules rather than their indirect expression from vectors.

Traditionally, key weaknesses of membrane-disruption strategies have been (1) the inconsistent level of cell-to-cell plasma membrane injury, with too little rendering insufficient delivery and too much causing excessive cell damage; (2) poor throughput and scalability (for example, microinjection); and (3) inadequate understanding of cell recovery, resulting in inefficient protocols<sup>47,48</sup>. Methods that employ severe thermal shock or electric fields may denature proteins or damage cell components<sup>49</sup>. Moreover, because of how membrane perturbation is administered in these techniques, they are often restricted to adherent or suspension cells. To overcome such challenges, new technologies are reinvigorating old approaches and concepts, suggesting that this may be a critical time in the development of *in vitro* and *ex vivo* intracellular delivery approaches (Box 1). Membrane disruption has seen promising advancement in recent years through nanotechnology, microfluidics, and laboratory-on-chip devices. Next, we cover the fundamentals of membrane disruption and repair before highlighting examples of technological progress in the field.

## The make and break of a cell membrane

The plasma membrane can be perturbed by physical means, with mechanical force, thermal deviations, electromagnetic radiation

and electric fields, or by appropriate biochemical agents, such as membrane-active peptides, detergents and pore-forming toxins<sup>50</sup> (Fig. 3a). A relatively straightforward mode is mechanical disruption, where in-plane tensile strains of 2%–3% rupture a lipid bilayer<sup>51</sup>. Such mechanical disruption can be administered through solid contact<sup>52–54</sup>, fluid shear<sup>55–57</sup>, or hydrostatic/osmotic pressure<sup>58</sup>. Depending on the contact area and strain rate the applied force may either disrupt the membrane immediately, or first deplete membrane reservoirs<sup>59</sup>. For example, sharp objects like microneedles concentrate the force to a small region and presumably penetrate rapidly, while the ‘blunt’ and relatively slower onset of an osmotic shock tends to deplete global reservoirs before producing disruptions<sup>59</sup>. Thermal deviations may promote membrane defects through several mechanisms. First, the higher kinetic energy associated with supraphysiological temperatures gives rise to more intense molecular fluctuations and subsequent dissociation of lipids<sup>60</sup>. In live cells it has been shown that leakage begins at 42 °C while temperatures above 55 °C promote rapid exchange of low-molecular-weight (<1 kDa) molecules<sup>60</sup>. Second, at the opposite end of the spectrum (<0 °C), formation of ice crystals can trigger mechanical expansion and cracking of the cell membrane, which may be repairable upon thawing<sup>61</sup>. Third, within the physiological range (0–40 °C), passing the membrane rapidly through thermal phase transitions may lead to the generation of holes, especially at phase domain boundaries<sup>60,61</sup>. In electroporation, the electric charging of non-conducting lipid bilayers and ever-present thermal fluctuations conspire to create and expand a heterogeneous population of pores<sup>62</sup>. As membrane potential rises beyond 0.2 V, hydrophobic defects of thermal origin readily transition over their energy barrier and expand into hydrophilic pores of 1 nm or more. Because small pores are poor conductors, their growth is energetically favourable while the field is maintained but slows down when they grow to a certain size<sup>62</sup>. Depending on the orientation of the electric field, these small pores may be unevenly distributed across the cell surface<sup>63</sup>. Alternatively, optoporation with lasers<sup>64</sup> produces a hole at a discrete point on the cell surface, and may involve a combination of mechanical, thermal and chemical effects, depending on pulse parameters<sup>65</sup>. Possibilities include thermal dissociation, thermoelastic stresses, effects beyond the focal region

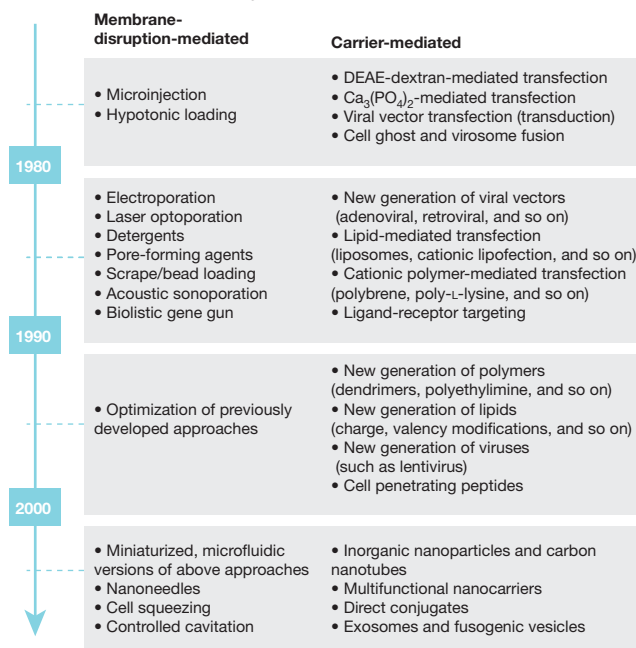
### BOX 1

## Snapshot of historical trends in intracellular delivery

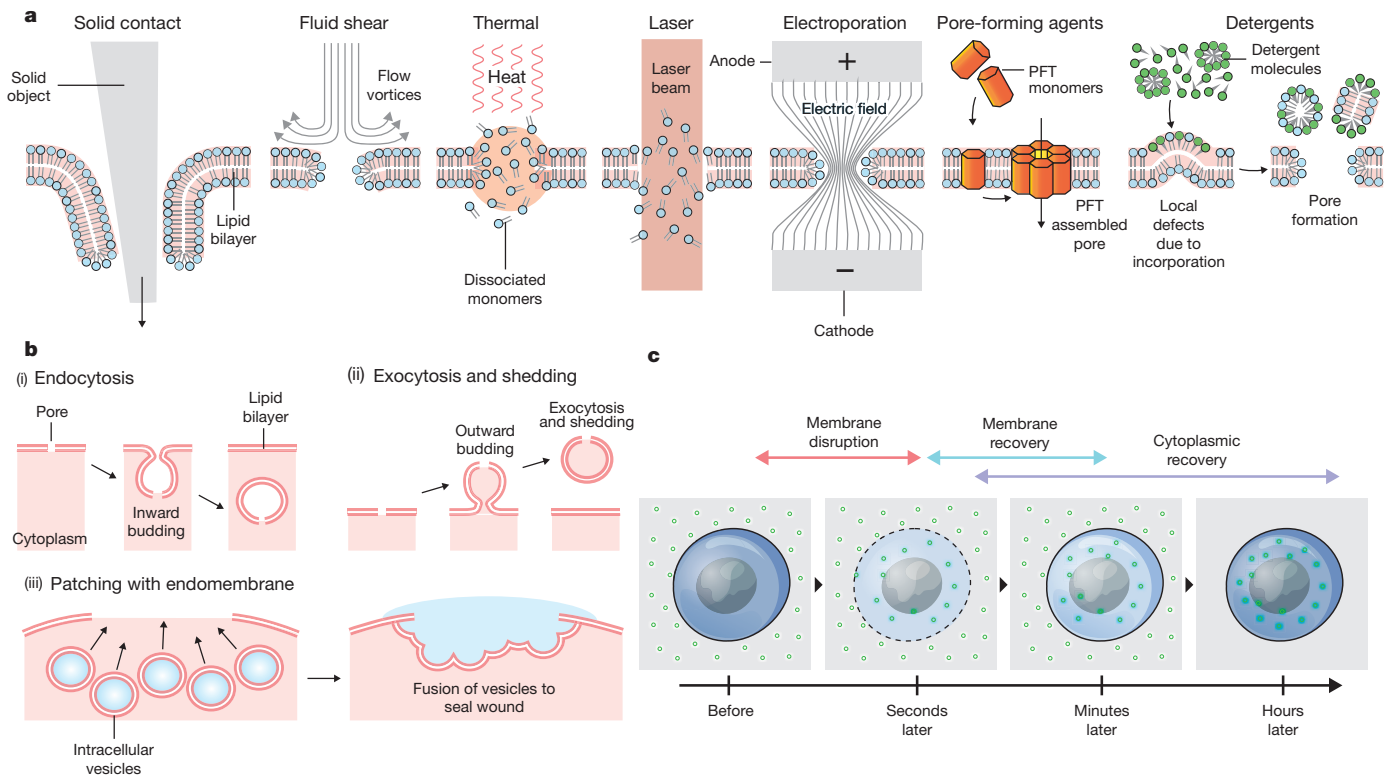
The field of intracellular delivery arguably began with the advent of microinjection<sup>52</sup> in 1911. Since then a broad range of options have evolved, which can be categorized into carrier-mediated or membrane-disruption-mediated (Fig. 2). The timeline highlights trends in the field.

For carrier-mediated approaches, early research noted that several cationic compounds readily complex with the negatively charged nucleic acids to facilitate uptake of DNA and RNA. Examples include precipitates formed with diethylaminoethyl (DEAE)-dextran and calcium phosphate. Inspired by these initial findings, chemical complexes and modified viruses were subsequently deployed as tools for DNA transfection. Since then hundreds of viral, lipid, polymer, and inorganic carriers have been developed, mostly for nucleic acid transfection<sup>1,24,30</sup>. Recent efforts have focused on multifunctional capabilities conferred through nanotechnology<sup>26,29</sup> and biomimetic strategies<sup>25</sup> such as exosomes<sup>40</sup>, direct conjugates<sup>2</sup>, and new generations of viral vectors<sup>22,23</sup>.

Membrane-disruption-based approaches have evolved in parallel. Initially, available options included low-throughput microinjection<sup>52</sup> or membrane perturbation with hypotonic shock<sup>58</sup>. The demonstration of DNA transfection by electroporation in 1982<sup>77</sup> sparked an era of widespread experimentation with other membrane-disruption modalities, including laser optoporation<sup>64</sup>, scrape/bead loading<sup>54</sup>, syringe loading<sup>55</sup>, acoustic sonoporation<sup>90</sup>, ballistic particles<sup>53</sup>, and permeabilization with detergents and pore-forming agents<sup>47,48</sup>. Electroporation rapidly gained a foothold as commercial products were launched from the mid-1980s. Most other membrane-disruption techniques were not broadly adopted, presumably owing to poor cell recoveries, limited throughput, need for specialized equipment, high cost, or skill-dependent operation<sup>47,48</sup>. In the past decade, nanotechnology, microfluidics, and laboratory-on-chip platforms are emerging as new possibilities to reinvigorate membrane-disruption-mediated approaches.







**Figure 3 | Membrane disruption and recovery in the context of intracellular delivery.** **a**, Schematic of plasma membrane disruption by mechanical forces (solid contact and fluid shear), thermal effects, focused lasers, electric fields, pore-forming agents that assemble into complexes, and the action of detergents that solubilize membrane lipids. **b**, Selected mechanisms of plasma membrane repair in the literature. Small disruptions (<100 nm) can be removed via endocytosis (i) and exocytosis or shedding of blebs (ii). Repair of large disruptions involves patching with

(such as shock-wave emission and shear stresses from induced cavitation bubbles) and generation of low-density free-electron plasma and reactive oxygen species. The latter leads to chemical degradation via peroxidation or reactive fragmentation of biomolecules<sup>65,66</sup>. Membrane defects from continuous-wave lasers are thought to arise from local heating, while nanosecond lasers produce a combination of heating, bubble formation and thermoelastic stresses. Femtosecond laser mechanisms are tunable on the basis of irradiance strength and repetition frequency, ranging from almost purely chemical effects to combinations of chemical and thermal degradation<sup>66</sup>.

Various biochemical agents have been used to permeabilize cells, the most eminent of which are detergents and pore-forming toxins. Pore-forming toxins approach the cell membrane as soluble agents, bind to the cell surface, oligomerize, and insert as an assembled pore complex<sup>67</sup>. The most often reported is Streptolysin O, owing to its ability to generate large pores of over 30 nm in diameter for the passage of proteins and large molecules<sup>68</sup>. Alternatively, detergents act by solubilizing membrane components. The amphiphilic plant glycosides digitonin and saponins are the most popular detergents for reversible permeabilization of cells<sup>47,48</sup>. Although the exact mechanisms are still a matter of study<sup>69</sup>, it is known that digitonin and saponin interact with membrane cholesterol, making them specific for the cholesterol-rich plasma membrane. However, the reported pore sizes are inconsistent, varying from a few nanometres up to the micrometre scale<sup>47,48,69</sup>.

Upon plasma membrane injury, the cell must either reseal or die. Along with the intended influx of cargo, there is the entry of  $\text{Ca}^{2+}$ , and the efflux of  $\text{K}^+$ , proteins, amino acids, metabolites, ATP, and other cytoplasmic contents to contend with<sup>70</sup>. Inundation with reactive oxygen species and other toxic molecules may cause further damage to endogenous proteins and biomolecules<sup>49</sup>. Thus, cells urgently deploy repair pathways to reseal

endomembrane from intracellular vesicles (contents in light blue) (iii). See recent reviews for further details<sup>70–74</sup>. **c**, Timescales of intracellular delivery via membrane disruption. Initially, membrane disruption permits the exchange of intracellular and extracellular contents, including cargo (green). Upon repair of membrane barrier function, delivered materials are retained while cellular recovery processes work to restore cytoplasmic homeostasis. PFT, pore-forming toxin.

membrane disruptions and recover from the damage imposed (see recent reviews<sup>70–74</sup>). This repair is an active process, primarily triggered by the influx of  $\text{Ca}^{2+}$  down its 10,000-fold concentration gradient<sup>75</sup>. There is a marked consistency in the properties of the repair pathways, regardless of whether the source of damage is electrical, mechanical, optical, or even chemical<sup>70,74</sup>. Instead, membrane recovery is thought to depend on disruption size, collateral damage, temperature, composition of the extracellular medium, and cell type. Up to six membrane repair pathways have been proposed, primarily involving membrane-trafficking processes around the defect<sup>71</sup>. As the exact mechanisms remain controversial<sup>70,71,73,74</sup>, we illustrate three broad concepts here (Fig. 3b). First,  $\text{Ca}^{2+}$ -dependent exocytosis and fusion of membrane-proximal vesicles leads to resealing by patch formation<sup>75</sup>, but may also serve to reduce tension around the wound, or release lysosomal signals for membrane remodelling<sup>74</sup>. Second and third, for smaller disruptions of several hundred nanometres or less, endocytosis or exocytosis, respectively, may extract the lesion into a disposable vesicle. The timescales of these repair processes are anywhere from a few seconds to several minutes.

Taken together, to achieve effective membrane-disruption-based delivery, the disruption must be sufficient to introduce the intended cargo, yet the cell must be capable of repairing itself without permanent damage. The membrane-disruption step usually occurs on a sub-second timescale; molecules then diffuse, or are driven, into the cell while repair and the associated contraction of holes takes place over seconds to minutes (Fig. 3c). This is followed by a longer phase involving restoration of cytoplasmic composition, stress response and possible alterations in transcription<sup>70</sup>. The influx and retention of molecules will depend on the size, lifetime and distribution of disruptions, as well as the properties of cargo molecules and their interaction with the cell<sup>50</sup>. Electroporation,

for example, provides an electrophoretic force that may boost the influx of certain charged molecules<sup>63,76</sup>.

### Towards precision membrane disruption

For decades, a leading delivery technique has been electroporation, with its ability to introduce diverse biomolecules to millions of cells per run. In a conventional setup, a solution with suspended cells is dispersed between parallel plates that apply a series of electrical pulses of determinable voltage, duration, waveform and frequency<sup>62,63,76,77</sup>. Compared to bulk electroporation setups, microfluidic designs offer the ability to localize the electric field to the scale of the cell (Fig. 4a). They can reduce the required voltage (often by 100-fold), provide superior heat dissipation, and incorporate flexible design features, such as hydrodynamic focusing to distance cells from potentially damaging electrodes<sup>78</sup>. An elegant and uncomplicated design was reported by ref. 79, who combined constant voltage with cells flowing through constrictions. The pulse strength was dictated by the cross-sectional ratio between the main channel and the constrictions, while the pulse duration was determined by speed of passage, thus avoiding the need for a pulse generator. Electroporation with a constant direct-current voltage has also been demonstrated for cells in aqueous droplets<sup>80</sup>. Owing to the non-conductivity of oil, cells only experience a transient electric pulse when the conductive droplets pass the electrodes. At the nanoscale, ref. 81 described a nanochannel electroporation system that uses a  $\sim 90$ -nm aperture to localize the permeabilization to a single point on the cell surface. This innovation enables the generation of a single large hole rather than the numerous small pores characteristic of conventional electroporation, which is less amenable to free passage of large materials. For example, conventional electroporation appears to exploit the charge of nucleic acids, such as plasmids and mRNA, in order to partially embed them in pores, resulting in subsequent internalization through active membrane-trafficking pathways rather than direct delivery<sup>81</sup>. In contrast, nanochannel electroporation achieves improved dose control, enhanced electrophoretic delivery deeper into the cell, and the ability to deliver materials that bulk electroporation often cannot, such as quantum dots.

Classic options for membrane disruption via mechanical perturbation include scrape and bead loading of adherent cells<sup>54</sup> or syringe loading of cells in suspension by repeated aspiration and expulsion through small gauge needles<sup>55</sup>. These methods provide coarse, inconsistent damage over a target cell population while being high throughput and low cost. On the other hand, modern microfabrication technology is now enabling mechanical approaches with improved precision (Fig. 4b, c). Prominent examples include cell squeezing, nanoneedles, and exploding cavitation bubbles. Cell squeezing involves the rapid deformation of cells as they passage through microfluidic constrictions of around half to one-third of the cell's diameter<sup>45</sup>. Diffusive delivery of a variety of cargoes including proteins, nucleic acids, quantum dots, carbon nanotubes and other nanomaterials has been demonstrated<sup>45</sup>. A major strength of squeezing is the simplicity of the device, with no moving parts or need for an external power supply. The energy for membrane disruption comes from the flow through a static structure. Although the technology has shown applicability across dozens of cell types at throughputs up to a million cells per second, a current limitation is the correlation between cell size and delivery efficiency. For an asynchronous population, cells that are too large may be lysed, such as multinuclear cells. Alternatively, cells that are too small do not experience sufficient deformation to disrupt the membrane properly. This has prompted the design of various constriction geometries to address different sizes of cell<sup>45</sup>.

Nanoneedles are another key development over the last decade, involving the generation of nanometre-scaled features capable of penetrating the cell membrane and providing access to the cytosol<sup>82</sup>. In the reported modalities thus far, the target material is delivered through one of three modes: (1) dissociation from the penetrating structure upon cytosolic entry<sup>82,83</sup>; (2) direct injection through a hollow nanoneedle or "nanostraw"<sup>84,85</sup> or (3) the diffusion from the extracellular medium through holes after withdrawal of the needles<sup>44</sup>. For the first mode, the

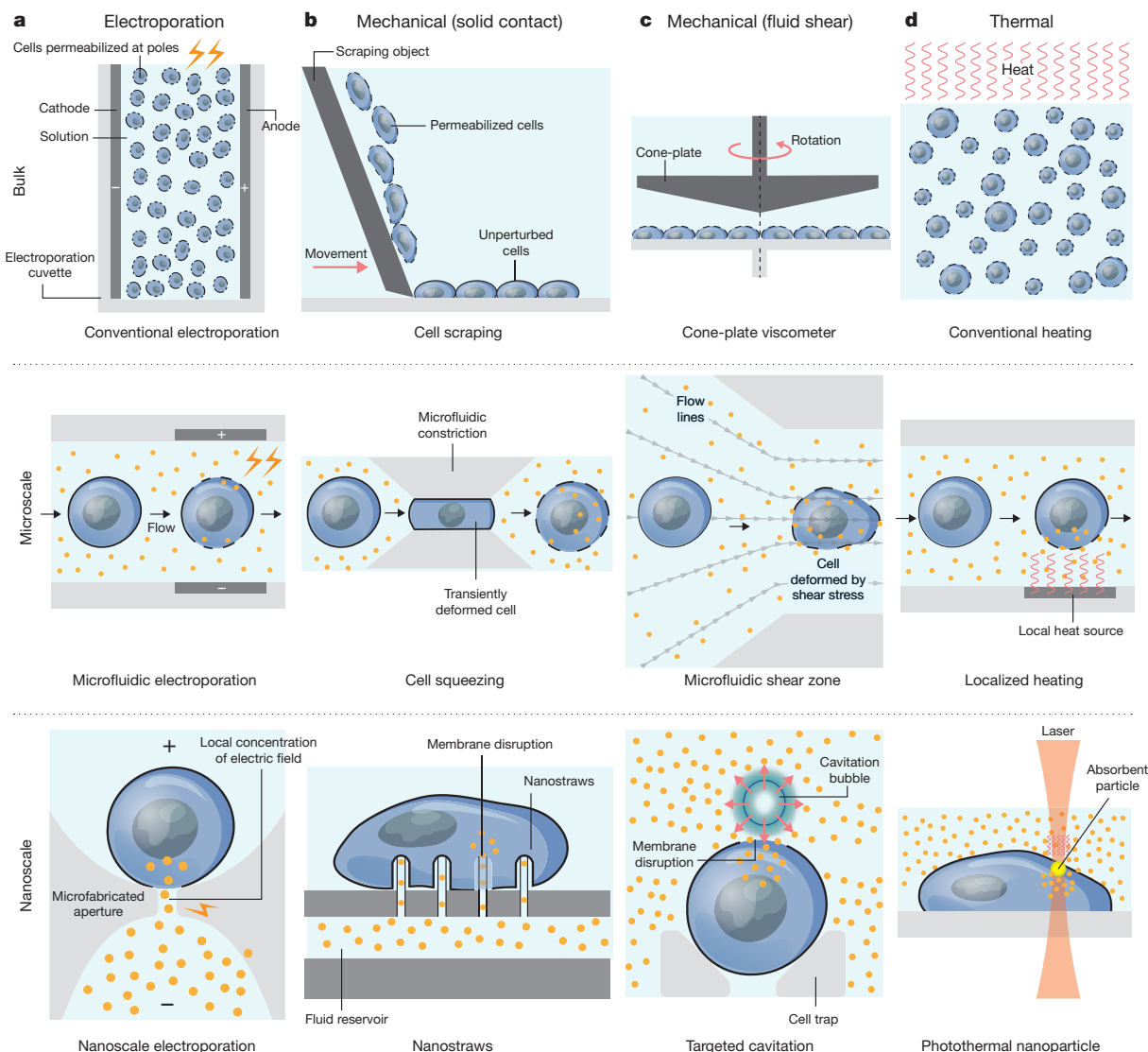
delivery of siRNA, peptides, DNA, proteins, and impermeable inhibitors to challenging cell types such as neurons and immune cells has been demonstrated<sup>83,86</sup>. Studies of the mechanism suggest that puncture does not occur upon initial cell contact, but requires active forces generated by cell spreading and formation of tension-promoting focal adhesions<sup>87</sup>. In the second mode, molecules have been successfully pumped into cells through nanostraws. A key benefit of this configuration is temporal control over delivery dynamics, volume, and dosage concentration, as well as possible gating with electric fields<sup>84,85</sup>. In the third mode, ref. 44 used a standard laboratory centrifuge to spin down a grid of diamond nanoneedles onto adherent cultured cells, followed by withdrawal and diffusive entry of cargo from solution. With this strategy it was possible to deliver a wide variety of cargo to primary neurons while maintaining  $>80\%$  viability. For nanoneedles of diameter about 300 nm and height of about 4  $\mu\text{m}$ , the required force for penetration was estimated to be 2 nN per needle. These mechanistic studies emphasize the need for active force to generate efficient membrane disruption. Thus far, the challenges of these systems have included the difficulty of fabricating high-precision structures and implementing the concept at scale.

Moving from solid to fluid (Fig. 4c)<sup>55</sup>, cone-plate viscometers, devices for generating a determinable shear force over a surface, have also been shown to transiently permeabilize apical membranes in cell monolayers<sup>56</sup>. Such observations inspired the design of microfluidic devices to expose cells to shear forces proportional to flow velocity through microchannels that taper from 300  $\mu\text{m}$  to 50  $\mu\text{m}$  in diameter<sup>57</sup>. Although an interesting concept, reproducibly controlling fluid shear forces at this scale has generally proven difficult. One way around this is to employ micrometre-sized cavitation bubbles<sup>88</sup>. Ultrasound phenomena, also known as sonoporation in the context of delivery, can produce cavitation bubbles in solution<sup>89</sup> and has been an intriguing permeabilization approach since its introduction in the 1980s<sup>90</sup>. However, a recent analysis across multiple published data sets indicates that ultrasound consistently struggles to deliver molecules with greater than 50% efficiency or 50% viability *in vitro*<sup>89</sup>. This was attributed to the operational mechanism of random and violent cavitation being heterogeneous, with some cells undergoing excessive damage while others remain unaffected. Targeted cavitation is a more promising idea, whereby a precisely positioned cavitation bubble is used to generate a local shear force at a given stand-off distance from a target cell<sup>91,92</sup>. Spatial control can be achieved by laser excitation of an absorbent particle or substrate. Recently, this concept was scaled up for deployment with cell monolayers<sup>93</sup>. Substrates arrayed with pores lined by metallic absorbers were irradiated underneath adherent cells. Exploding bubbles were synchronized with active pumping to successfully introduce large cargo, such as living bacteria greater than one micrometre, into the cytoplasm of several cell types.

Bulk thermal insults are capable of perturbing the membrane for delivery of small molecules but detrimental effects on cell function preclude their implementation<sup>60,61</sup>. Thermal disruption may be more feasible if confined to a localized area (Fig. 4d). Indeed, gene transfection has been demonstrated with cell solutions processed through thermal inkjet printers<sup>94</sup>, although it is unclear whether membrane perturbation of passing cells arises from fluid shear effects at the nozzle, temperature spikes, or both. A more precise approach is the use of absorbent nanoparticles as nucleation sites for intense local heating<sup>65</sup>. Upon laser irradiation local perturbation effects may be due to thermal, chemical, or cavitation-mediated fluid shear phenomena. Recent efforts indicate that tuning the laser pulse parameters and properties of the absorbing particles can bias the mechanism towards a particular mode<sup>95,96</sup>. Although parameters are still being explored, a proof-of-concept study demonstrated conditions under which irradiation of gold nanoparticles was proposed to trigger localized thermal damage that permitted more than half of the treated cells to take up labelled antibodies<sup>97</sup>.

As illustrated above, advances in micro- and nanotechnology are breathing new life into delivery modes that were once deemed impractical (Fig. 4). By concentrating precise membrane-perturbing effects to the





**Figure 4 | Selected modes of bulk, microscale and nanoscale approaches for membrane-disruption-based intracellular delivery.** Molecular cargo is shown in orange.

cellular and subcellular scale, the potential exists to address applications that are underserved by current techniques. Early progress suggests that such techniques may be among the first real challengers to electroporation's long-held dominance on membrane-disruption-based intracellular delivery. Nanoneedles and microfluidics, for example, have shown promising compatibility in stem cell reprogramming<sup>45</sup>, functional interrogation of primary immune cells<sup>86</sup>, and transfection of cultured neurons<sup>44</sup>. To achieve further advances, delivery efficacy and safety must be combined with scalability, tunable throughput, low cost and user-friendliness.

## Outlook

Every day in research institutes and clinical centres around the world, scientists use kits and protocols based on viral vectors, lipid transfection agents, and electroporation, among other options. The complex mechanisms of established methods and their often unpredictable impact on cell behaviour have dramatically limited the scope of biological experiments and reduced efficacy of potentially promising cell therapy concepts. The biomedical research community would benefit greatly from a more mechanistic and transparent understanding of intracellular delivery, both to further the development of more robust techniques and to realize key medical and industrial applications.

Effective and safe intracellular delivery systems facilitate progress in multiple fields from cell-based therapies, gene editing and cutting-edge

genomics to reprogramming cellular states and probing the intracellular environment (Fig. 1). Demand for effective solutions currently outstrips supply by a large margin, however. Nowhere is this aberration more greatly felt than in the treatment of primary and patient-derived cells, including various types of immune cells, neurons and stem cells. Deep interdisciplinary coordination will be required to transform imaginative engineering solutions into technological platforms with biological compatibility and relevance. Mechanistic studies must seek to gauge delivery performance quantitatively, and to assess influx mechanisms, cell damage and off-target perturbations associated with treatment. More rigorous analysis of these factors will benefit the field, and counter the temptation to 'overhype' technologies that are yet to prove their utility in the clinic or laboratory<sup>98,99</sup>. Issues related to scale-up, cost and compatibility with current goods and manufacturing protocols must be considered at an earlier stage of development. To this end, we anticipate that the guidelines in Tables 1 and 2 will assist researchers in selecting appropriate methods and help aspiring inventors to understand key requirements and areas of opportunity.

As emerging intracellular delivery technologies take us beyond routine nucleic acid transfection and enable robust manipulation of previously recalcitrant cell types, we will be entering exciting new territory. The ability to deliver proteins, peptides, nanomaterials, molecular tags, and a variety of other compounds will enable unprecedented flexibility

in our capacity to manipulate cell function and probe the intracellular environment. The systematic deployment of transcription factors to alter gene expression, or antibodies to label or block intracellular processes, could be revolutionary. With next-generation intracellular delivery, the development of molecular probes and nanomaterial sensors with which to analyse intracellular properties could facilitate both discovery and diagnostics with unprecedented accuracy. Future approaches could focus on targeting specific subcellular compartments—developing combinatorial strategies to direct materials to the endoplasmic reticulum, nucleus or mitochondria, for example<sup>100</sup>.

Advances in nanotechnology and microfluidics will continue to expand the frontiers of membrane-disruption-based delivery. Already, platforms harnessing the power of exploding bubbles, microfluidic squeezing, and nanoneedles have been transformed into commercial ventures. On the other hand, carrier-based technologies continue to develop, with new generations of viral vectors, endosome disruption strategies, stimuli-sensitive functional materials, and biomimetic inspiration from pathogenic mechanisms and membrane-trafficking processes. Progress in these fields will lead to new challenges, such as off-target effects and innate cellular responses against carrier and cargo alike. Solving this next generation of problems may hinge on our ability to understand current delivery mechanisms and to implement the analytical approaches necessary to characterize cellular responses. Despite the barriers that remain, we anticipate that next-generation technologies will translate beyond academic endeavours into portable, personalized, cell-based diagnostics and the use of clinical intracellular delivery to engineer cell fate for therapeutic benefit.

Received 11 March 2015; accepted 11 July 2016.

1. Yin, H. *et al.* Non-viral vectors for gene-based therapy. *Nat. Rev. Genet.* **15**, 541–555 (2014).
  2. Wittrup, A. & Lieberman, J. Knocking down disease: a progress report on siRNA therapeutics. *Nat. Rev. Genet.* **16**, 543–552 (2015).
  3. Naldini, L. *Ex vivo* gene transfer and correction for cell-based therapies. *Nat. Rev. Genet.* **12**, 301–315 (2011).
  4. Naldini, L. Gene therapy returns to centre stage. *Nature* **526**, 351–360 (2015).
  5. June, C. H., Riddell, S. R. & Schumacher, T. N. Adoptive cellular therapy: a race to the finish line. *Sci. Transl. Med.* **7**, 280ps7 (2015).
  6. Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* **348**, 62–68 (2015).
  7. Tebas, P. *et al.* Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. *N. Engl. J. Med.* **370**, 901–910 (2014).
  8. Sather, B. D. *et al.* Efficient modification of CCR5 in primary human hematopoietic cells using a megaTAL nuclease and AAV donor template. *Sci. Transl. Med.* **7**, 307ra156 (2015).
  9. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
  10. Cox, D. B., Platt, R. J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nat. Med.* **21**, 121–131 (2015).
  11. Kim, D. *et al.* Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell Stem Cell* **4**, 472–476 (2009).
  12. Warren, L. *et al.* Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell* **7**, 618–630 (2010).
  13. Anokye-Danso, F. *et al.* Highly efficient miRNA-mediated reprogramming of mouse and human somatic cells to pluripotency. *Cell Stem Cell* **8**, 376–388 (2011).
  14. Marschall, A. L. J. *et al.* Delivery of antibodies to the cytosol: debunking the myths. *Mabs* **6**, 943–956 (2014).
  15. Liu, J., Wen, J., Zhang, Z., Liu, H. & Sun, Y. Voyage inside the cell: microsystems and nanoengineering for intracellular measurement and manipulation. *Microsyst. Nanoeng.* **1**, 15020 (2015).
  16. Lee, S. E., Liu, G. L., Kim, F. & Lee, L. P. Remote optical switch for localized and selective control of gene interference. *Nano Lett.* **9**, 562–570 (2009).
  17. Heller, D. A., Baik, S., Eurell, T. E. & Strano, M. S. Single-walled carbon nanotube spectroscopy in live cells: Towards long-term labels and optical sensors. *Adv. Mater.* **17**, 2793–2799 (2005).
  18. Michalet, X. *et al.* Quantum dots for live cells, *in vivo* imaging, and diagnostics. *Science* **307**, 538–544 (2005).
  19. Karra, D. & Dahm, R. Transfection techniques for neuronal cells. *J. Neurosci.* **30**, 6171–6177 (2010).
  20. Peer, D. A daunting task: manipulating leukocyte function with RNAi. *Immunol. Rev.* **253**, 185–197 (2013).
  21. Hendel, A. *et al.* Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* **33**, 985–989 (2015).
  22. Thomas, C. E., Ehrhardt, A. & Kay, M. A. Progress and problems with the use of viral vectors for gene therapy. *Nat. Rev. Genet.* **4**, 346–358 (2003).
  23. Kay, M. A. State-of-the-art gene-based therapies: the road ahead. *Nat. Rev. Genet.* **12**, 316–328 (2011).
  24. Mintzer, M. A. & Simanek, E. E. Nonviral vectors for gene delivery. *Chem. Rev.* **109**, 259–302 (2009).
  25. Yoo, J. W., Irvine, D. J., Discher, D. E. & Mitragotri, S. Bio-inspired, bioengineered and biomimetic drug delivery carriers. *Nat. Rev. Drug Discov.* **10**, 521–535 (2011).
  26. Torchilin, V. P. Multifunctional, stimuli-sensitive nanoparticulate systems for drug delivery. *Nat. Rev. Drug Discov.* **13**, 813–827 (2014).
  27. Mitragotri, S., Burke, P. A. & Langer, R. Overcoming the challenges in administering biopharmaceuticals: formulation and delivery strategies. *Nat. Rev. Drug Discov.* **13**, 655–672 (2014).
  28. Khalil, I. A., Kogure, K., Akita, H. & Harashima, H. Uptake pathways and subsequent intracellular trafficking in nonviral gene delivery. *Pharmacol. Rev.* **58**, 32–45 (2006).
  29. Sahay, G., Alakhova, D. Y. & Kabanov, A. V. Endocytosis of nanomedicines. *J. Control. Release* **145**, 182–195 (2010).
  30. Stewart, M. P., Lorenz, A., Dahlman, J. & Sahay, G. Challenges in carrier-mediated intracellular delivery: moving beyond endosomal barriers. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.* **8**, 465–478 (2016).
  31. Sahay, G. *et al.* Efficiency of siRNA delivery by lipid nanoparticles is limited by endocytic recycling. *Nat. Biotechnol.* **31**, 653–658 (2013).
  32. Gilleron, J. *et al.* Image-based analysis of lipid nanoparticle-mediated siRNA delivery, intracellular trafficking and endosomal escape. *Nat. Biotechnol.* **31**, 638–646 (2013).
  33. Schaffer, D. V., Fidelman, N. A., Dan, N. & Lauffenburger, D. A. Vector unpacking as a potential barrier for receptor-mediated polyplex gene delivery. *Biotechnol. Bioeng.* **67**, 598–606 (2000).
  34. Lv, H., Zhang, S., Wang, B., Cui, S. & Yan, J. Toxicity of cationic lipids and cationic polymers in gene delivery. *J. Control. Release* **114**, 100–109 (2006).
  35. Yang, B. *et al.* High-throughput screening identifies small molecules that enhance the pharmacological effects of oligonucleotides. *Nucleic Acids Res.* **43**, 1987–1996 (2015).
  36. Furusawa, M., Nishimura, T., Yamaizumi, M. & Okada, Y. Injection of foreign substances into single cells by cell fusion. *Nature* **249**, 449–450 (1974).
  37. Helenius, A., Doxsey, S. & Mellman, I. Viruses as tools in drug delivery. *Ann. NY Acad. Sci.* **507**, 1–6 (1987).
  38. Daemen, T. *et al.* Virosomes for antigen and DNA delivery. *Adv. Drug Deliv. Rev.* **57**, 451–463 (2005).
  39. Montecalvo, A. *et al.* Mechanism of transfer of functional microRNAs between mouse dendritic cells via exosomes. *Blood* **119**, 756–766 (2012).
  40. El Andaloussi, S., Mäger, I., Breakefield, X. O. & Wood, M. J. A. Extracellular vesicles: biology and emerging therapeutic opportunities. *Nat. Rev. Drug Discov.* **12**, 347–357 (2013).
  41. He, W. *et al.* Discovery of siRNA lipid nanoparticles to transfect suspension leukemia cells and provide *in vivo* delivery capability. *Mol. Ther.* **22**, 359–370 (2014).
  42. Blanco, E., Shen, H. & Ferrari, M. Principles of nanoparticle design for overcoming biological barriers to drug delivery. *Nat. Biotechnol.* **33**, 941–951 (2015).
  43. Van Meirvenne, S. *et al.* Efficient genetic modification of murine dendritic cells by electroporation with mRNA. *Cancer Gene Ther.* **9**, 787–797 (2002).
  44. Wang, Y. *et al.* Poking cells for efficient vector-free intracellular delivery. *Nat. Commun.* **5**, 4466 (2014).
  45. Sharei, A. *et al.* A vector-free microfluidic platform for intracellular delivery. *Proc. Natl Acad. Sci. USA* **110**, 2082–2087 (2013).
- In this paper, rapid mechanical deformation of cells through microfluidic constrictions was shown to achieve efficient intracellular delivery of a wide range of molecular cargo.**
46. Schumann, K. *et al.* Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. *Proc. Natl Acad. Sci. USA* **112**, 10437–10442 (2015).
  47. Schulz, I. Permeabilizing cells: some methods and applications for the study of intracellular processes. *Methods Enzymol.* **192**, 280–300 (1990).
  48. Hapala, I. Breaking the barrier: methods for reversible permeabilization of cellular membranes. *Crit. Rev. Biotechnol.* **17**, 105–122 (1997).
  49. Agarwal, J., Walsh, A. & Lee, R. C. Multimodal strategies for resuscitating injured cells. *Ann. NY Acad. Sci.* **1066**, 295–309 (2005).
  50. Gurtovenko, A. A., Anwar, J. & Vattulainen, I. Defect-mediated trafficking across cell membranes: insights from *in silico* modeling. *Chem. Rev.* **110**, 6077–6103 (2010).
  51. Bloom, M., Evans, E. & Mouritsen, O. G. Physical properties of the fluid lipid-bilayer component of cell membranes: a perspective. *Q. Rev. Biophys.* **24**, 293–397 (1991).
  52. Barber, M. A. A technic for the inoculation of bacteria and other substances into living cells. *J. Infect. Dis.* **8**, 348–360 (1911).
- This is arguably amongst the first reports on intracellular delivery, demonstrating basic microinjection of isolated cells.**
53. Klein, T. M., Wolf, E. D., Wu, R. & Sanford, J. C. High-velocity microprojectiles for delivering nucleic-acids into living cells. *Nature* **327**, 70–73 (1987).
- This report introduced the gene gun for intracellular delivery by exploiting biolistic particles.**
54. Mcneil, P. L. Incorporation of macromolecules into living cells. *Methods Cell Biol.* **29**, 153–173 (1988).



55. Clarke, M. S. F. & McNeil, P. L. Syringe loading introduces macromolecules into living mammalian cell cytosol. *J. Cell Sci.* **102**, 533–541 (1992).
56. LaPlaca, M. C., Lee, V. M. Y. & Thibault, L. E. An *in vitro* model of traumatic neuronal injury: loading rate-dependent changes in acute cytosolic calcium and lactate dehydrogenase release. *J. Neurotrauma* **14**, 355–368 (1997).
57. Hallow, D. M. *et al.* Shear-induced intracellular loading of cells with molecules by controlled microfluidics. *Biotechnol. Bioeng.* **99**, 846–854 (2008).
58. Borle, A. B. & Snowdowne, K. W. Measurement of intracellular free calcium in monkey kidney cells with aequorin. *Science* **217**, 252–254 (1982).
59. Groulx, N., Boudreault, F., Orlov, S. N. & Grygorczyk, R. Membrane reserves and hypotonic cell swelling. *J. Membr. Biol.* **214**, 43–56 (2006).
60. Bischof, J. C. *et al.* Dynamics of cell membrane permeability changes at supraphysiological temperatures. *Biophys. J.* **68**, 2608–2614 (1995).
61. He, X. M., Amin, A. A., Fowler, A. & Toner, M. Thermally induced introduction of trehalose into primary rat hepatocytes. *Cell Preserv. Technol.* **4**, 178–187 (2006).
62. Weaver, J. C. & Chizmadzhev, Y. A. Theory of electroporation: a review. *Bioelectrochem. Bioenerg.* **41**, 135–160 (1996).
63. Kandußer, M. & Miklavčič, D. in *Electrotechnologies for Extraction from Food Plants and Biomaterials* (eds Vorobiev, E. & Lebovka, N.) Ch. 1, 1–37 (Springer, 2008).
64. Tsukakoshi, M., Kurata, S., Nomiya, Y., Ikawa, Y. & Kasuya, T. A novel method of DNA transfection by laser microbeam cell surgery. *Appl. Phys. B* **35**, 135–140 (1984).
- This pioneering paper demonstrated DNA transfection of mammalian cells by laser optoporation.**
65. Stevenson, D. J., Gunn-Moore, F. J., Campbell, P. & Dholakia, K. Single cell optical transfection. *J. R. Soc. Interf.* **7**, 863–871 (2010).
66. Vogel, A., Noack, J., Huttman, G. & Palttauf, G. Mechanisms of femtosecond laser nanosurgery of cells and tissues. *Appl. Phys. B* **81**, 1015–1047 (2005).
- This publication outlines a comprehensive framework covering the mechanisms of laser–membrane interactions.**
67. Bischofberger, M., Iacovache, I. & van der Goot, F. G. Pathogenic pore-forming proteins: function and host response. *Cell Host Microbe* **12**, 266–275 (2012).
68. Walev, I. *et al.* Delivery of proteins into living cells by reversible membrane permeabilization with streptolysin-O. *Proc. Natl Acad. Sci. USA* **98**, 3185–3190 (2001).
69. Frenkel, N., Makky, A., Sudji, I. R., Wink, M. & Tanaka, M. Mechanistic investigation of interactions between steroidal saponin digitonin and cell membrane models. *J. Phys. Chem. B* **118**, 14632–14639 (2014).
70. Cooper, S. T. & McNeil, P. L. Membrane repair: mechanisms and pathophysiology. *Physiol. Rev.* **95**, 1205–1240 (2015).
71. Moe, A. M., Golding, A. E. & Bement, W. M. Cell healing: calcium, repair and regeneration. *Semin. Cell Dev. Biol.* **45**, 18–23 (2015).
72. Jimenez, A. J. & Perez, F. Physico-chemical and biological considerations for membrane wound evolution and repair in animal cells. *Semin. Cell Dev. Biol.* **45**, 2–9 (2015).
73. Boucher, E. & Mandato, C. A. Plasma membrane and cytoskeleton dynamics during single-cell wound healing. *Biochim. Biophys. Acta* **1853**, 2649–2661 (2015).
74. Andrews, N. W., Corrotte, M. & Castro-Gomes, T. Above the fray: surface remodeling by secreted lysosomal enzymes leads to endocytosis-mediated plasma membrane repair. *Semin. Cell Dev. Biol.* **45**, 10–17 (2015).
75. Steinhart, R. A., Bi, G. & Alderton, J. M. Cell membrane resealing by a vesicular mechanism similar to neurotransmitter release. *Science* **263**, 390–393 (1994).
- This groundbreaking paper alerted the research community to the active and responsive nature of cell plasma membrane repair, which was previously attributed to passive resealing.**
76. Venslauskas, M. S. & Šatkauskas, S. Mechanisms of transfer of bioactive molecules through the cell membrane by electroporation. *Eur. Biophys. J.* **44**, 277–289 (2015).
77. Neumann, E., Schaefer-Ridder, M., Wang, Y. & Hofschneider, P. H. Gene transfer into mouse lymphoma cells by electroporation in high electric fields. *EMBO J.* **1**, 841–845 (1982).
- This pioneering paper demonstrated DNA transfection of mammalian cells by electroporation.**
78. Movahed, S. & Li, D. Q. Microfluidics cell electroporation. *Microfluidics Nanofluidics* **10**, 703–734 (2011).
79. Geng, T. *et al.* Flow-through electroporation based on constant voltage for large-volume transfection of cells. *J. Control. Release* **144**, 91–100 (2010).
80. Zhan, Y., Wang, J., Bao, N. & Lu, C. Electroporation of cells in microfluidic droplets. *Anal. Chem.* **81**, 2027–2031 (2009).
81. Boukany, P. E. *et al.* Nanochannel electroporation delivers precise amounts of biomolecules into living cells. *Nat. Nanotechnol.* **6**, 747–754 (2011).
82. McKnight, T. E. *et al.* Intracellular integration of synthetic nanostructures with viable cells for controlled biochemical manipulation. *Nanotechnology* **14**, 551–556 (2003).
- This pioneering paper constitutes the first major implementation of nanoneedle arrays for DNA transfection.**
83. Shalek, A. K. *et al.* Vertical silicon nanowires as a universal platform for delivering biomolecules into living cells. *Proc. Natl Acad. Sci. USA* **107**, 1870–1875 (2010).
- This study expanded the use of nanoneedles for co-delivery of diverse biomolecules.**
84. VanDersarl, J. J., Xu, A. M. & Melosh, N. A. Nanostraws for direct fluidic intracellular access. *Nano Lett.* **12**, 3881–3886 (2012).
85. Xie, X. *et al.* Nanostraw-electroporation system for highly efficient intracellular delivery and transfection. *ACS Nano* **7**, 4351–4358 (2013).
86. Shalek, A. K. *et al.* Nanowire-mediated delivery enables functional interrogation of primary immune cells: application to the analysis of chronic lymphocytic leukemia. *Nano Lett.* **12**, 6498–6504 (2012).
87. Xu, A. M. *et al.* Quantification of nanowire penetration into living cells. *Nat. Commun.* **5**, 3613 (2014).
88. Marmottant, P. & Hilgenfeldt, S. Controlled vesicle deformation and lysis by single oscillating bubbles. *Nature* **423**, 153–156 (2003).
89. Liu, Y., Yan, J. & Prausnitz, M. R. Can ultrasound enable efficient intracellular uptake of molecules? A retrospective literature review and analysis. *Ultrasound Med. Biol.* **38**, 876–888 (2012).
90. Fecshheimer, M. *et al.* Transfection of mammalian cells with plasmid DNA by scrape loading and sonication loading. *Proc. Natl Acad. Sci. USA* **84**, 8463–8467 (1987).
- This pioneering paper demonstrated DNA transfection of mammalian cells by ultrasound.**
91. Prentice, P., Cuschier, A., Dholakia, K., Prausnitz, M. & Campbell, P. Membrane disruption by optically controlled microbubble cavitation. *Nat. Phys.* **1**, 107–110 (2005).
92. Ohl, C. D. *et al.* Sonoporation from jetting cavitation bubbles. *Biophys. J.* **91**, 4285–4295 (2006).
93. Wu, Y. C. *et al.* Massively parallel delivery of large cargo into mammalian cells with light pulses. *Nat. Methods* **12**, 439–444 (2015).
94. Cui, X., Dean, D., Ruggeri, Z. M. & Boland, T. Cell damage evaluation of thermal inkjet printed Chinese hamster ovary cells. *Biotechnol. Bioeng.* **106**, 963–969 (2010).
95. Xiong, R. *et al.* Comparison of gold nanoparticle mediated photoporation: vapor nanobubbles outperform direct heating for delivering macromolecules in live cells. *ACS Nano* **8**, 6288–6296 (2014).
96. Schomaker, M. *et al.* Characterization of nanoparticle mediated laser transfection by femtosecond laser pulses for applications in molecular medicine. *J. Nanobiotechnol.* **13**:10, <http://dx.doi.org/10.1186/s12951-014-0057-1> (2015).
97. Yao, C., Qu, X., Zhang, Z., Hüttmann, G. & Rahmzadeh, R. Influence of laser parameters on nanoparticle-induced membrane permeabilization. *J. Biomed. Opt.* **14**, 054034 (2009).
98. Time to deliver. *Nat. Biotechnol.* **32**, 961 (2014).
99. Park, K. Facing the truth about nanotechnology in drug delivery. *ACS Nano* **7**, 7442–7447 (2013).
100. Rajendran, L., Knölker, H. J. & Simons, K. Subcellular targeting strategies for drug design and delivery. *Nat. Rev. Drug Discov.* **9**, 29–42 (2010).

**Acknowledgements** This work was supported by the US National Institute of Health (R01GM101420-01A1). M.P.S. was supported by the Swiss NSF through the advanced postdoc mobility fellowship P300P3\_151179. M.P.S. acknowledges support from a Keith Murdoch Fellowship via the American Australian Association, a Life Sciences Research Foundation Fellowship sponsored by Good Ventures, and a Broadnext10 Catalytic Steps funding gift from the Broad Institute. A.S. was supported by a Ragon Institute fellowship. We thank the following people for comments and constructive criticism: E. van Leen, D. Irvine, J. Voldman, S. Manalis, J. Weaver, J. Lieberman, R. Karnik, R. Lee, D. Mueller, S. Bhakdi, Y. Toyoda, Z. Maliga, H.-J. Lee, N. Yang, E. Lim, R. Sayde and K. Blagovic.

**Author Contributions** R.L. and K.F.J. shaped ideas and provided guidance. M.P.S. constructed the figures. M.P.S. and A.S. wrote the manuscript (with assistance from X.D. and G.S.).

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.L. ([rlanger@mit.edu](mailto:rlanger@mit.edu)) or K.F.J. ([kfjensen@mit.edu](mailto:kfjensen@mit.edu)).

**Reviewer Information** *Nature* thanks L. Lee, M. Prausnitz and the other anonymous reviewer(s) for their contribution to the peer review of this work.

# The evolution of Ebola virus: Insights from the 2013–2016 epidemic

Edward C. Holmes<sup>1</sup>, Gytis Dudas<sup>2,3</sup>, Andrew Rambaut<sup>3,4,5</sup> & Kristian G. Andersen<sup>6,7,8</sup>

**The 2013–2016 epidemic of Ebola virus disease in West Africa was of unprecedented magnitude and changed our perspective on this lethal but sporadically emerging virus. This outbreak also marked the beginning of large-scale real-time molecular epidemiology. Here, we show how evolutionary analyses of Ebola virus genome sequences provided key insights into virus origins, evolution and spread during the epidemic. We provide basic scientists, epidemiologists, medical practitioners and other outbreak responders with an enhanced understanding of the utility and limitations of pathogen genomic sequencing. This will be crucially important in our attempts to track and control future infectious disease outbreaks.**

The 2013–2016 Ebola virus disease (EVD) epidemic in West Africa appears to have begun following human contact with an animal (probably bat) reservoir of Ebola virus (EBOV) in December 2013, in the small village of Meliandou in Guéckédou Prefecture, Guinea<sup>1</sup>. After this initial spill-over infection, the outbreak remained undetected for several months and spread via chains of sustained human-to-human transmission, with no evidence of additional zoonotic transfers from the animal reservoir<sup>1–4</sup>. By the time that EBOV (a lineage later named the Makona variant<sup>5</sup>) was confirmed in March 2014, several villages, towns and larger cities had reported cases<sup>1</sup>. When the World Health Organization declared the EVD outbreak to constitute a Public Health Emergency of International Concern in August 2014<sup>6</sup>, EBOV had already spread across country borders with more than a thousand cases reported in Guinea, Sierra Leone, Liberia and Nigeria. In the epidemic that followed, a total of 28,646 confirmed and suspected cases of EVD were documented, with 11,323 recorded deaths, making it by far the largest outbreak of EVD on record<sup>7</sup>.

Ebola virus is a negative-sense single-strand RNA ((–)ssRNA) virus with a 19-kilobase genome and, like most other RNA viruses, quickly generates mutations through error-prone replication. Until recently, genomic studies of infectious disease outbreaks were necessarily retrospective, occurring after the pathogen had either been eradicated or developed endemic transmission in the host population<sup>8–12</sup>. However, recent developments in high-throughput next-generation sequencing (NGS)<sup>13–16</sup> enabled rapid and in-depth viral genomic surveillance during the 2013–2016 EVD epidemic<sup>1–3,17–26</sup>. Indeed, with the advent of NGS it is now possible to generate pathogen genomic data directly from diagnostic patient samples<sup>2,3,17–27</sup> within days or hours of the sample being taken<sup>25,26</sup>, and in challenging field situations<sup>19,23,25,26</sup>. The resulting large-scale sequence data sets provide new opportunities for the epidemiological investigation of transmission chains and the improvement of outbreak responses<sup>28</sup>. In the case of the 2013–2016 EVD epidemic, the sequence data generated have revealed key aspects of the patterns and processes of EBOV evolution as the epidemic proceeded<sup>2,3,20–22,24–26,29,30</sup>. Hence, not only was the 2013–2016 epidemic a landmark in the epidemiological history of EBOV, but the size of the resulting genomic data set—over 1,500 full-length EBOV Makona sequences (Table 1), or approximately 5% of those

infected—also makes it one of the most densely sampled infectious disease outbreaks (Fig. 1a, b). Although sequence data have been generated from outbreaks of viral disease for over 30 years<sup>8–12,31,32</sup>, the sheer size of the

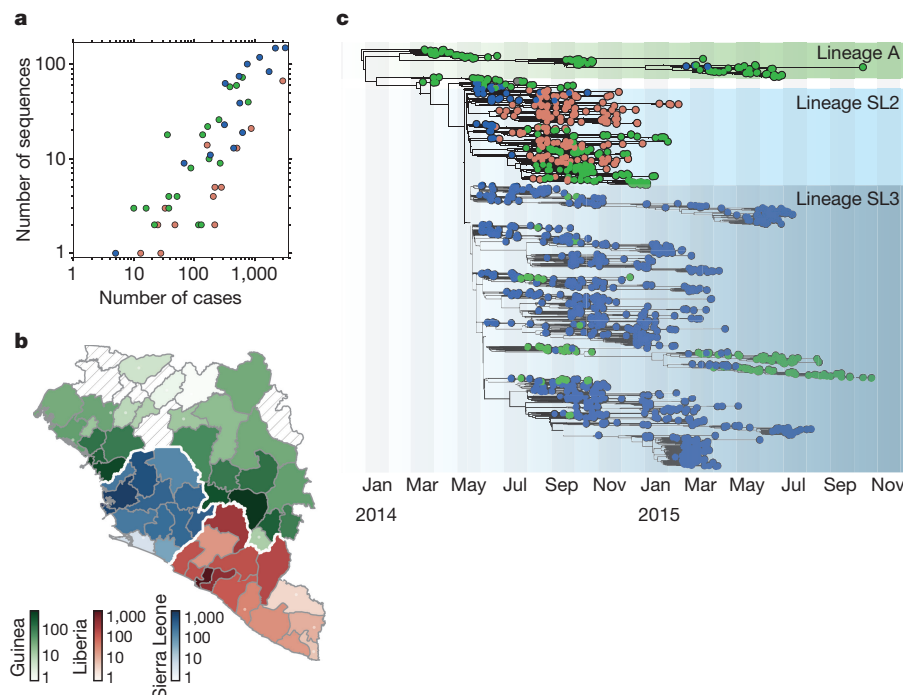
**Table 1 | Overview of EBOV sequencing studies performed during the 2013–2016 epidemic**

Study	Platform	Method	Sequencing location	Case location	No. of seqs
Baize, S. <i>et al.</i> (Apr. 2014) <sup>1</sup>	Sanger	Amplified	International	Guinea	3
Gire, S. K. <i>et al.</i> (Sep. 2014) <sup>2</sup>	Illumina	Direct	International	Sierra Leone	79
Hoenen, T. <i>et al.</i> (Apr. 2015) <sup>17</sup>	Sanger	Amplified	International	Mali	4
Bell, A. <i>et al.</i> (May. 2015) <sup>18</sup>	Illumina	Direct	International	UK	3
Park, D. J. <i>et al.</i> (Jun. 2015) <sup>3</sup>	Illumina	Direct	International	Sierra Leone	232
Kugelman, J. R. <i>et al.</i> (Jul. 2015) <sup>19</sup>	Illumina	Direct	In-country/ Liberia	Liberia	25
Simon-Loriere, E. <i>et al.</i> (Aug. 2015) <sup>20</sup>	Illumina	Direct	International	Guinea	85
Carroll, M. W. <i>et al.</i> (Aug. 2015) <sup>21</sup>	Illumina	Direct	International	Guinea/ Liberia	179
Tong, Y. G. <i>et al.</i> (Aug. 2015) <sup>22</sup>	BGISEQ-100	Amplified	?	Sierra Leone	175
Smits, S. L. <i>et al.</i> (Sep. 2015) <sup>23</sup>	Ion Torrent	Amplified	In-country/ Sierra Leone	Sierra Leone	49
Ladner, J. T. <i>et al.</i> (Dec. 2015) <sup>24</sup>	Illumina	Direct	International	Liberia	140
Quick, J. <i>et al.</i> (Feb. 2016) <sup>25</sup>	MinION	Amplified	In-country/ Guinea	Guinea	137
Hoenen, T. <i>et al.</i> (Feb. 2016) <sup>93</sup>	MinION	Amplified	In-country/ Liberia	Liberia	8
Arias, A. <i>et al.</i> (Jun. 2016) <sup>26</sup>	Ion Torrent	Amplified	In-country/ Sierra Leone	Sierra Leone	554

Summary of the different sequencing efforts performed during the 2013–2016 EVD epidemic, noting sequencing platform. We include the following parameters for each study: method (Direct, no PCR amplification and/or enrichment; Amplified, material amplified via amplicon-based PCR before viral sequencing), sequencing location, country of origin for the sequenced samples (case location) and number of EBOV genomes produced.

<sup>1</sup>Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and Sydney Medical School, Charles Perkins Centre, University of Sydney, Sydney, New South Wales 2006, Australia. <sup>2</sup>Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. <sup>3</sup>Institute of Evolutionary Biology, Ashworth Laboratories, University of Edinburgh, Edinburgh EH9 3FL, UK. <sup>4</sup>Centre for Immunology, Infection and Evolution, University of Edinburgh, Ashworth Laboratories, Edinburgh EH9 3FL, UK. <sup>5</sup>Fogarty International Center, National Institutes of Health, MSC 2220 Bethesda, Maryland 20892, USA. <sup>6</sup>The Scripps Research Institute, Department of Immunology and Microbial Science, La Jolla, San Diego, California 92037, USA. <sup>7</sup>The Scripps Research Institute, Department of Integrative Structural and Computational Biology, La Jolla, San Diego, California 92037, USA. <sup>8</sup>Scripps Translational Science Institute, La Jolla, San Diego, California 92037, USA.





**Figure 1 | Evolution of EBOV during the 2013–2016 outbreak showing the extent and location of virus sampling.** **a**, Sampling during the 2013–2016 EVD epidemic. Sequencing efforts closely match confirmed and suspected case numbers in each administrative division of Guinea (green), Liberia (red) and Sierra Leone (blue) (Spearman correlation coefficient = 0.91). **b**, Map of the three countries most affected by EVD during the 2013–2016 EVD epidemic. Administrative divisions in Guinea, Liberia and Sierra Leone are shown in green, red and blue, respectively, and coloured according to the cumulative numbers of confirmed and suspected cases throughout the epidemic. Hatched areas indicate divisions that never reported any cases. The boundary data for the maps is from GADM (<http://www.gadm.org>). **c**, Temporal phylogeny of all publicly available EBOV genomes estimated using BEAST<sup>97</sup>. Three lineages

identified in previous studies<sup>2,21,25</sup> are marked with coloured backgrounds. The sequence alignment was partitioned into four categories: codon positions 1, 2 and 3, and non-coding intergenic regions. Changes in each of the four partitions were modelled according to the HKY+ $\Gamma_4$  nucleotide substitution model with relative rates between partitions. Tip dates were used to calibrate a relaxed molecular clock with rates drawn from a lognormal distribution<sup>54</sup> with an uninformative prior placed on the mean of the distribution. A flexible 'skygrid' tree prior was used to allow for changes in effective population sizes over time. Each tip is coloured according to the country where the patient was most likely to have been infected: green for Guinea, red for Liberia and blue for Sierra Leone. Data correct as of 19 April, 2016. *Nature* remains neutral with regard to jurisdictional claims in published maps.

data set, the widespread spatial coverage (Fig. 1a), and the contemporary nature of the EBOV data provide the first in-depth genomic anatomy of an epidemic, setting a benchmark for future outbreak responses. Here, we describe how pathogen sequences produced during the 2013–2016 EVD epidemic provided key insights into EBOV genomic epidemiology and molecular evolution, and note the lessons that need to be learned for the effective study of future outbreaks.

### Ebola virus disease in humans

Ebola virus (species *Zaire ebolavirus*) is one of four viruses—with Sudan virus, Tai Forest virus, and Bundibugyo virus—within the genus *Ebolavirus* that cause severe disease in humans and other primates. The final member of the genus is Reston virus, although infection with this virus does not appear to cause human disease<sup>33</sup>. All ebolaviruses are members of the family *Filoviridae*, which also includes *Lloviu cuevavirus* (genus *Cuevavirus*) and the severe human pathogen Marburg virus (genus *Marburgvirus*). It is believed that bats serve as the primary reservoir for EBOV<sup>34,35</sup>. However, EBOV infections have been confirmed in only a small number of mammalian species and it is unclear whether the virus infects a wider range of animal hosts that have yet to be sampled. Some evidence for a broader host distribution, at least in the evolutionary past, comes from the observation that endogenous filoviruses are present in the genomes of diverse mammalian species, including marsupials<sup>36,37</sup>.

EBOV in humans was first described in Zaire (now the Democratic Republic of the Congo (DRC)) in 1976, where, over a two-month period, it led to an outbreak of 318 cases with an 88% case-fatality rate (CFR)<sup>38</sup>. CFRs, however, are difficult to estimate for EVD<sup>39,40</sup>, so such numbers

should be interpreted with caution. Between 1977 and 2014, 12 smaller outbreaks were reported in Middle Africa, with 32–315 cases and CFRs ranging from 47% to 89%<sup>41</sup>. The 2013–2016 EVD epidemic is therefore notable not only for its duration and magnitude, but also as the first outbreak in West Africa and the first in which case exportations and nosocomial transmissions were reported outside of Africa<sup>41</sup>. However, despite the scale of the 2013–2016 EVD epidemic, infection with EBOV Makona appears to lead to similar disease characteristics and transmission profiles as previous EBOV outbreak variants<sup>42,43</sup>. For example, the CFR for the 2013–2016 EVD epidemic appears to be around 70%<sup>1,39,40,43</sup> and estimates for the basic reproduction number ( $R_0$ ) fall between 1.5 and 2.5, both of which are comparable to calculations from previous outbreaks<sup>39,44–47</sup>.

### Origin of the 2013–2016 Ebola virus disease epidemic

Evolutionary analyses of genome sequences from the 2013–2016 EVD epidemic have provided a clear picture of the origin and spread of EBOV Makona<sup>1–3,20–22,24–26,48</sup>. One of the most important early questions was whether the epidemic was the result of a single cross-species transmission event into humans, or whether there were repeated zoonotic events from a widespread animal EBOV reservoir. Owing to the high genetic similarity of virus genomes sampled from the beginning of the epidemic, a single spill-over infection seems the more likely<sup>1,2</sup>. Phylogenetic analyses also make it clear that once the outbreak was established, later lineages of EBOV Makona had descended from those circulating earlier in the epidemic<sup>2,3,20–22,24,25</sup> (Fig. 1c). This is in contrast to some of the earlier EVD outbreaks, in which epidemiological and sequence-based investigations have provided evidence for multiple spill-over infections<sup>49,50</sup>.

Sequence-based findings are consistent with epidemiological investigations into the timing of the 2013–2016 EVD epidemic, which placed the first case around late December 2013 in Guinea<sup>1</sup>. In agreement with this, molecular clock dating analyses suggest that the common ancestor of all sequenced EBOV Makona lineages be placed at the beginning of 2014<sup>2,21,48</sup>, with lineages in Guinea falling close to the root of the tree (Fig. 1c). These studies also showed that EBOV Makona diverged from other EBOV outbreak variants only about a decade ago<sup>2,48</sup>. This finding suggests that EBOV Makona may be fairly new to West Africa, sharing recent common ancestry with Middle African variants that are found thousands of miles away. Molecular clock dating analyses have also shown that all recorded human EVD outbreaks caused by EBOV appear to share a common ancestor around 1975<sup>2,51</sup>. Notably, this is around the time of the first described EVD outbreak in 1976, suggesting that the EBOV lineage experienced a severe genetic bottleneck before the first human outbreak<sup>52,53</sup>. Despite their power<sup>54</sup>, molecular clock dating studies of this type would undoubtedly benefit from additional EBOV genomic sequence data from both previous EVD outbreaks and animal reservoir populations.

### Genetic diversification of Ebola virus Makona

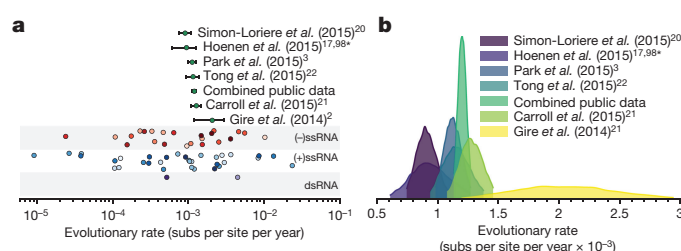
Because of the relatively small magnitude and duration of previous EVD outbreaks, earlier EBOV sequencing efforts were necessarily limited to small numbers of temporally clustered cases. The data from these earlier studies largely comprised single viral lineages and led to the perception that EBOV genomes remain stable over the course of an outbreak<sup>55–59</sup>. However, the much larger size and duration of the 2013–2016 EVD epidemic (Fig. 1a, b) resulted in a different molecular epidemiological pattern for EBOV Makona, in which multiple virus lineages arose and co-circulated (Fig. 1c).

Despite their shared border, the EBOV Makona genomes sampled from the three most affected countries, Guinea, Sierra Leone and Liberia, generally (although not exclusively) form separate clusters on phylogenetic trees and exhibit different phylogenetic patterns<sup>3,19–22,24–26</sup> (Fig. 1b, c). Genomic studies have shown that the 2013–2016 EVD epidemic was dominated by three major lineages, denoted A (refs 21, 25), SL2 (ref. 2) and SL3 (refs 2, 3) (Fig. 1c). Most of these lineages—including lineage A<sup>21,25</sup> in Guinea, SL3 in Sierra Leone<sup>3</sup> and Liberian isolates<sup>24</sup>—circulated locally, with only sporadic cross-border transmission (Fig. 1c). By contrast, lineage SL2 (ref. 2) was the most widespread in the region<sup>3,21,22,24,25</sup> (Fig. 1c). This lineage probably arose in Sierra Leone<sup>2</sup> where it gave rise to lineage SL3 and several sub-lineages<sup>3,22</sup>. It crossed more than twice into Liberia<sup>24</sup>, seeded several transmission chains in Guinea<sup>21</sup> and spread throughout Sierra Leone<sup>2,3,22</sup> (Fig. 1c). It is unclear whether any of these lineages carry mutations that could have affected their epidemic potential<sup>60</sup>, or, perhaps more likely, whether the increased geographical spread of SL2 and SL3 is a reflection of chance epidemiological founding events (see below)<sup>60,61</sup>.

### Evolutionary dynamics of Ebola virus Makona

Although the origin and spread of the 2013–2016 EVD epidemic seem well resolved<sup>1–3,20–22,24,25</sup>, other aspects of EBOV evolution during this epidemic have proven more controversial. A major point of contention in both scientific publications<sup>2,3,20–22,24,62–64</sup> and the popular press<sup>65,66</sup> has been whether the virus ‘mutated’ unusually rapidly during this outbreak. Unfortunately, much of this discussion is based on misrepresentations of what type of rate was measured and how these rates can be translated into predictions of phenotypic evolution.

The starting point for the debate over how quickly EBOV Makona evolved was the observation by Gire *et al.*<sup>2</sup> that the mean evolutionary rate early in the epidemic was  $\sim 1.9 \times 10^{-3}$  (95% Bayesian credible interval:  $1.11, 2.91 \times 10^{-3}$ ) nucleotide substitutions (subs) per site per year. This rate was approximately twice as high as that averaged from genomic sequences of EBOV variants sampled from multiple outbreaks, at around  $0.9 \times 10^{-3}$  ( $0.81, 1.18 \times 10^{-3}$ ) subs per site per year (ref. 2). EBOV outbreak variants are viral lineages responsible for human outbreaks. Other EBOV variants include EBOV Yambuku (Mayinga) from



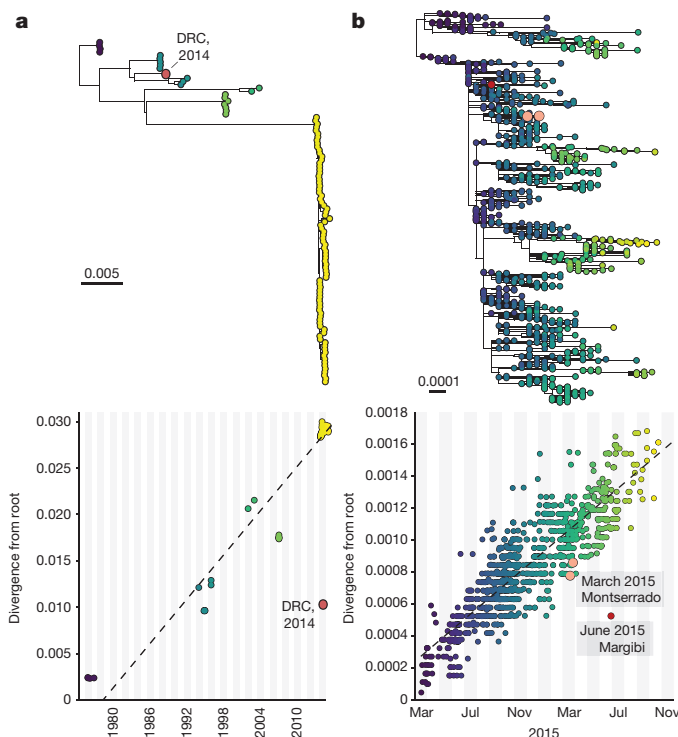
**Figure 2 | Evolutionary rates of EBOV compared to those of other RNA viruses.** **a**, Estimates of evolutionary rate in diverse RNA viruses. Green points at the top indicate the mean evolutionary rates estimated for EBOV during the 2013–2016 EVD epidemic from different studies, with solid lines showing the 95% credible intervals derived from BEAST analyses. Points at the bottom represent equivalent estimates (without uncertainty intervals) published previously for negative-sense single-strand RNA viruses (red), positive-sense single-strand RNA viruses (blue) and double-strand RNA viruses (purple)<sup>72</sup>. Points with the same shade belong to the same family. Evolutionary rate estimates for EBOV Makona occupy a narrow distribution within the range of rates observed in RNA viruses as a whole. **b**, 95% credible intervals for the distribution of evolutionary rates for EBOV from the 2013–2016 EVD epidemic published previously. \*An erratum<sup>98</sup> revised the mean evolutionary rate estimate for ref. 17 (Hoenen *et al.*), to  $1.32 \times 10^{-3}$  (95% credible intervals:  $0.89, 1.75 \times 10^{-3}$ ) subs per site per year.

1976, EBOV Kikwit from 1995, and EBOV Lomela from 2014 (DRC). The between-outbreak evolutionary rate therefore reflects estimates averaged across all EBOV variants. However, later studies of EBOV Makona consistently produced lower rate estimates than those generated by Gire *et al.*<sup>3,21,22,25</sup>. Indeed, taking the publicly available sequence data as a whole, estimates of the EBOV evolutionary rate for the 2013–2016 epidemic converge on a mean value of around  $1.2 \times 10^{-3}$  ( $1.13, 1.27 \times 10^{-3}$ ) (Fig. 2). The ensuing discussions of whether EBOV is evolving more or less rapidly than expected, and what this means for the ability of the virus to evolve changes in transmissibility and virulence, have become a common narrative<sup>67–70</sup>.

The debate over the evolutionary dynamics of EBOV highlights a number of general issues in viral evolution. First, estimates of evolutionary rates are generally expected to be higher within outbreaks than between them. This is because the relatively short timescale over which sequences are sampled during outbreaks may be insufficient for mutations to be removed (or make them less likely to be fixed) by either natural selection or genetic drift. Hence, pathogen genomic sequences sampled early in epidemics will contain an excess of mildly deleterious variants that would eventually be eliminated by purifying selection<sup>61</sup>. This will tend to inflate evolutionary rates and in part explains why evolutionary rates in RNA viruses are often ‘time-dependent’: high towards the present, low towards the past<sup>71,72</sup>. Indeed, it is notable that as the 2013–2016 EVD epidemic progressed, analyses of evolutionary rate in EBOV converged on a reliable estimate (Fig. 2), which is expected owing to the increasing size of the data set combined with a longer sampling period. When viewed in the context of viruses as a whole, it is also striking that all the evolutionary rate estimates for EBOV fall in a narrow range towards the centre of a distribution that spans more than three orders of magnitude, from about  $10^{-2}$  to about  $10^{-5}$  subs per site per year (Fig. 2).

As well as time-dependence, it is possible that purifying selection may be relaxed in humans following cross-species transmission and/or that EBOV may undergo more replications per unit time during human outbreaks than in its reservoir species<sup>2,73</sup>. Both of these scenarios would increase the within-outbreak rate. Potential evidence for fundamental differences in evolutionary dynamics associated with species-jumping is provided by the EBOV Lomela variant that emerged in the DRC during 2014, causing a small EVD outbreak with 69 cases<sup>74</sup>. The branch length on the EBOV phylogenetic tree leading to the EBOV Lomela sequences from their common ancestor is far shorter than expected from their sampling time in 2014 (Fig. 3a), indicating a markedly lower evolutionary rate<sup>74,75</sup>.





**Figure 3 | Examples of violations of the Ebola virus molecular clock.**

**a**, Root-to-tip regression of genetic distances against time (month and year) of sampling for 105 representative EBOV variant sequences collected between 1976 and 2016 based on a maximum likelihood tree. **b** Equivalent root-to-tip regression of publicly available sequences from the 2013–2016 EVD epidemic using data on the day of sampling, and the maximum likelihood tree on which the estimates were made. RAXML (ref. 99) (panel **a**) and PhyML (ref. 100) (panel **b**) were used to estimate the maximum likelihood phylogenies under an HKY+ $\Gamma_4$  substitution model that was rooted via least squares regression in TempEst. Substitutions accumulate linearly with time, with some variation. Sequences recovered from transmission events that occurred as a result of persistent EBOV infection often exhibit temporal anomalies. In this scenario, EBOV may accumulate substitutions at a lower rate during persistence in individuals compared to regular person-to-person transmission. Larger red points indicate sequences of EBOV sampled from EVD survivor-associated transmission chains<sup>76,77</sup>. Scale bar, nucleotide substitutions per site.

This could reflect an evolutionary history in a different reservoir host from those previously described for EBOV, in which replication rates and hence evolutionary rates are reduced, or in which purifying selection acts with greater potency than in humans. Lower EBOV evolutionary rates were also observed in suspected cases of transmission from human EVD survivors during the 2013–2016 epidemic (Fig. 3b)<sup>76,77</sup>. Unexpectedly low evolutionary rates may therefore serve as an important signal for detecting probable transmissions from EVD survivors during flare-ups<sup>26,76</sup> (Fig. 3b).

The debate over EBOV evolutionary rate estimates has also revealed confusion over the terminologies used to describe the rate at which genetic changes accumulate. The most straight-forward measure of the rate of molecular evolution is the nucleotide substitution rate. This rate describes the frequency with which mutations are fixed in populations through time and for EBOV is best approximated by the rate observed between outbreaks<sup>61</sup> (Box 1). This rate reflects the long-term evolutionary processes including selective constraints on the genome, host-species-specific adaptation and the cumulative results of genetic drift. In contrast, the rate of change within outbreaks might be better thought of as the evolutionary rate, as the short timescale of sampling necessarily means that not all mutations observed will be fixed. Both the substitution rate and evolutionary rate can be clearly distinguished from the mutation rate. This term relates to the rate at which mutations are generated during viral

replication by intrinsic biochemical factors, and in particular to how frequently the viral polymerase makes errors<sup>78</sup> (Box 1, Box 1 Figure). This rate is generally challenging to measure<sup>79,80</sup> and is unknown for most viruses, including EBOV. It is therefore unfortunate that the debate over EBOV evolution has focused on ‘mutation’ (and hence potential differences intrinsic to particular virus lineages) when this is not the parameter that has been measured.

Finally, it is too simplistic to think that a twofold variation in rate estimates for EBOV will result in radically different evolutionary behaviour, especially when seen in the context of RNA viruses as a whole (Fig. 2). The likelihood of meaningful adaptive evolution depends not only on the rate at which the virus is able to generate mutations, but also on those environmental and host factors that shape the selection pressures acting on the virus. That filoviruses have infected a wide range of mammalian hosts<sup>36,37,81</sup> suggests that they are readily able to adapt to new environments irrespective of potential differences in evolutionary rate.

### Phenotypic evolution of Ebola virus Makona

While the patterns of EBOV molecular evolution during the 2013–2016 EVD epidemic have been well characterized, it is not currently known whether any of the observed mutations have resulted in differences in viral phenotype. This is particularly the case with respect to such traits as antigenicity, transmissibility and virulence, or mutations that could have an impact on vaccines, therapeutics and diagnostics. Although genomic sequence data play a central role in understanding outbreak dynamics and evolution, revealing key aspects of viral phenotype using sequence data alone is fraught with difficulties, and may even be counterproductive to outbreak response by steering the focus away from more critical needs<sup>65</sup>.

As the 2013–2016 West African epidemic of EVD was so much larger than previous outbreaks, it is possible that EBOV Makona possessed mutations that enhanced its transmissibility in humans. Without direct experimental data, however, a simpler scenario is that the scale and severity of the 2013–2016 EVD epidemic reflects a different epidemiological context than previous outbreaks. Under this model, most, if not all, EBOV variants entering human populations after cross-species transmission have the ability to cause major epidemics, but have been unable to do so because of a lack of a susceptible host population and/or environment. In particular, previous EVD outbreaks occurred in largely isolated and rural areas<sup>41</sup> (with the notable exception of the 1995 outbreak in Kikwit, which has a population of ~400,000; ref. 41), where there were either an insufficient number of susceptible people to guarantee long-term transmission, or the outbreak was quickly controlled by efficient interventions. The 2013–2016 EVD epidemic, in contrast, was the first in West Africa and the first in which a large EVD epidemic resulted in sustained community transmission from rural settings to major urban centres, where it was easier to establish large-scale transmission networks. This included the establishment of ‘underground’ networks, amplified by reluctance to seek medical advice in the affected communities, which greatly hindered intervention strategies focused on breaking chains of transmission. That the scale of the 2013–2016 EVD epidemic more reflects virus epidemiology rather than virus evolution is also supported by the failure to find evidence for heritable changes in the duration of virus shedding or virulence during the course of the 2013–2016 EVD epidemic<sup>39,40,42–47</sup>.

However, it was also the case that EBOV evolution during the 2013–2016 EVD epidemic was characterized by an abundance of changes in the nucleotide and amino acid sequences that could fuel adaptation for more efficient human transmission; any mutations that increased  $R_0$  would have been favoured by natural selection. Because of its key role in virus–host interactions, most attention has been directed towards the EBOV glycoprotein, and it is notable that the highest level of genetic amino acid diversity generated during the 2013–2016 EVD epidemic occurred in the glycoprotein (in particular, its mucin-like domain)<sup>3,22</sup>. For example, we observed 104 amino acid changes in glycoproteins that were shared by at least two EBOV Makona lineages from the 1,500 available EBOV genomes that make up 5%

## BOX 1

## Different measures of genome sequence change

**Mutation rate**

As viruses replicate, mutational errors are incorporated into the viral genome. The mutation rate is therefore typically expressed as the number of mutations per site, per replication event. The mutation rate for RNA viruses such as EBOV is largely determined by the viral RNA-dependent RNA polymerase, which lacks proof-reading activity. The estimation of mutation rates requires complex sequencing-based or phenotypic marker experiments that correct for the impact of natural selection<sup>79</sup>. Mutation rates are unknown for most viruses and have not been determined for EBOV, although it would be predicted to be comparable to other (–)ssRNA viruses and is likely to be similar across all EBOV outbreak variants.

**Evolutionary rate**

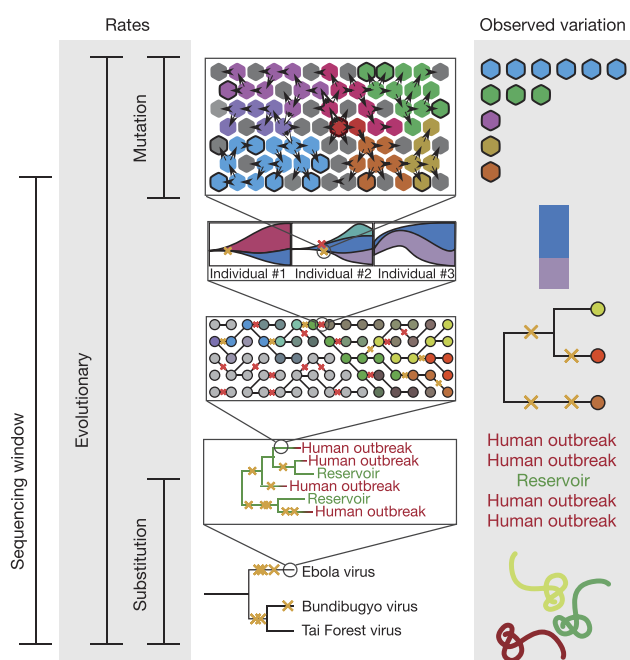
The evolutionary rate of a virus can be defined as the observed rate at which new variants arise and spread in the viral population. This can be measured by methods that compare the genetic change in viral genomes collected at different times. Importantly, evolutionary rates in RNA viruses may be dependent on the timescale over which they are measured: they are elevated in the short-term, such as within disease outbreaks, because mildly deleterious mutations may not have been eliminated by purifying selection<sup>71,72</sup>.

**Substitution rate**

The substitution rate is best described as the long-term rate at which genetic variants become fixed in a virus lineage over evolutionary time-scales, such as between human outbreaks in the case of EBOV. Hence, this rate reflects the complex interplay of natural selection, genetic drift, modes of transmission and epidemiological processes. This rate will also usually be lower than the short-term evolutionary rate because many of the variants circulating within outbreaks and epidemics will ultimately be eliminated. Furthermore, saturation—repeated changes at the same site—will further reduce the measured substitution rate.

**Fixation rate**

An added complexity in estimating rates in RNA viruses is that the population genetic concept of ‘fixation’, central to the definition of substitution, is ill-defined. In slowly evolving organisms, fixation events can usually be distinguished from polymorphisms by analysing individual nucleotide sites within and between species. However, in rapidly evolving RNA viruses, fixation can be described to occur (1) at the level of individual hosts over the course of infection, (2) in viral lineages within specific geographic locations or epidemiological networks (such as the different lineages of EBOV generated during the 2013–2016 EVD epidemic), (3) in global meta-populations, and (4) between different viral species.



**Box 1 Figure | Illustration of different measures of genomic variation.** Mutations accumulate over time. This phenomenon is at the core of molecular clocks, a class of methods that aim to convert molecular phylogenies with branch lengths given in expected substitutions per site into plausible temporal phylogenies in which branch lengths are given in time units and the trees themselves are embedded in time. By making use of sequences sampled at different times, such methods can estimate the evolutionary rate that provides the conversion from genetic distance into time. As phylogenetic methods have become ever more powerful and easily accessible, confusion has resulted from the frequent and interchangeable use of the terms mutation rate and substitution rate to signify the ‘molecular clock’ rate. Mutation and substitution rates, however, sit on the opposite ends of the evolutionary rate continuum and neither is the appropriate term for the molecular clock rate derived from densely sequenced epidemics.

of the more than 28,000 reported EVD cases<sup>7</sup>. While it is not known whether any of these amino acid changes led to functional differences, one plausibly important glycoprotein variant that originated early in the epidemic (in lineage SL2)<sup>24</sup> is an alanine to valine change at residue 82 (A82V). This is the first substitution observed in the receptor binding domain of EBOV and could potentially alter the interaction between the EBOV glycoprotein and its host receptor Niemann–Pick C1 (NPC1)<sup>82,83</sup>. Clearly, determining whether lineages of EBOV Makona

carrying A82V or other mutations that arose during the 2013–2016 EVD epidemic differ in epidemic potential should be a research priority.

Irrespective of potential differences in transmissibility that are yet to be uncovered, it is more certain that EBOV Makona is no different from previous EBOV outbreak variants when it comes to bodily fluids being the primary route of transmission<sup>81</sup>. Early on in the 2013–2016 EVD epidemic there was high-profile speculation that EBOV could evolve



respiratory (that is, airborne) transmission due to genetic diversity in the viral population<sup>62,64,65</sup>. However, there is no evidence for airborne EBOV transmission during the 2013–2016 EVD epidemic—or any other EVD outbreaks—and nor are there any examples of other viruses evolving a new mode of transmission on the timescale of individual outbreaks. Although influenza virus shifts its mode of transmission from (primarily) faecal–oral in its wild bird reservoir to respiratory in humans<sup>84</sup>, this change occurs at the point of cross-species transmission and not during human outbreaks.

While the occurrence of airborne transmission can be eliminated for EBOV, studies using genomic sequence data have conclusively shown that sexual transmission plays a previously unappreciated role for EBOV dissemination and reignition<sup>77,85–88</sup>. However, the long-term epidemiological and evolutionary implications of this mode of transmission are unclear and warrant further in-depth studies.

### Public health implications of genomic epidemiology

In addition to providing essential information on the pattern and dynamics of viral evolution during epidemics, viral genomic data may be of more direct public health importance. Indeed, the 2013–2016 EVD epidemic is arguably the first in which genomic data have been used directly in a real-time public health setting, to inform policies and infection control<sup>2,7,25,26</sup>. That some of these studies were undertaken under difficult field conditions<sup>19,23,25,26</sup> highlights the potential for portable genomic sequencing to transform outbreak responses<sup>7,25</sup>.

The simplest use of genomic data during outbreaks has been to reveal the pathways of viral spread through communities; when combined with phylogeographic approaches<sup>22,89</sup>, the results can be used to direct intervention methods to transmission hot-spots and to determine the impact of specific interventions such as border closures. For example, Tong and colleagues used viral genome sequencing to show how EBOV spread from the capital city of Freetown to multiple districts throughout Sierra Leone<sup>22</sup>, with Arias *et al.* later documenting how virus traffic from Freetown established new transmission clusters late in the epidemic<sup>26</sup>. Similarly, phylogenetic analyses revealed the co-circulation of multiple EBOV lineages within individual localities such as Conakry<sup>20</sup>, as well as cross-border virus traffic between Guinea and Sierra Leone<sup>25</sup>, highlighting important gaps in intervention. On a more localized epidemiological scale, genome sequence data provide a way to reveal who infected whom in EBOV transmission networks (although see below). Pathogen sequence data can therefore yield key information on the likelihood of, for example, sexual transmission<sup>77,85–88</sup>, as well as the possible transmission of EBOV via breast milk<sup>26</sup>. A similarly precise reconstruction of transmission chains is essential in understanding the multiple reignition events that occurred during the EBOV epidemic and their relation to viral transmissions from EVD survivors<sup>76</sup> (Fig. 3b). It is unclear whether the small subset of EVD survivors that harbour persistent infections pose a sustained infection risk or whether an episode of renewed viral replication is required for transmission to occur. Considering the pattern and degree of EBOV genetic change within such cases may provide critical insights. Phylogenetic approaches also provide a powerful way to accurately estimate various outbreak parameters, such as  $R_0$ , including that for individual virus lineages that are slow and difficult to obtain using longitudinal case data<sup>90,91</sup>. Finally, virus ‘super spreaders’ within human populations can also be readily identified using pathogen sequence data<sup>92</sup>.

Despite the quantity and quality of the viral genome sequence data generated during the 2013–2016 EVD epidemic, there are limitations to the scope and impact of genomic epidemiology. Clearly, the direct phenotypic effects of individual mutations on vaccines, therapeutics and diagnostics need to be tested experimentally. However, should viral lineages that differ in such properties arise during outbreaks, evolutionary genomic analyses can provide a powerful means to both determine their origins and rapidly track their spread through human populations.

### Lessons learned and future directions

The 2013–2016 EVD epidemic has set the benchmark for the use of large-scale molecular epidemiology as an essential tool in outbreak response.

Given the development of portable sequencing technologies, real-time viral genome sequencing is now possible in clinics and diagnostic laboratories, including in resource-limited settings<sup>25,93</sup>. This will offer critical data to inform epidemiological intervention, but will require a willingness to invest in scientific infrastructure, healthcare and training of local staff in the affected countries<sup>94</sup>. The need for immediate analysis and the growth of open sharing of sequence data means the challenge in genomic studies may be moving from data acquisition to analysis and interpretation. However, it is also the case that in-country real-time sequencing was not established until relatively late in the West African epidemic<sup>19,23,25,26,93</sup>, when case numbers had already begun to decline. In addition, many of the genome sequences were obtained in the absence of strong clinical and epidemiological metadata, such as the precise geographical location from where the sample was obtained, whether the individual survived the infection, and the time to the onset of symptoms. While it may be difficult to obtain such data during a rapidly developing outbreak, this limits the usefulness of genomic sequencing data in addressing a number of central biological questions, such as the virological basis to any variation in disease presentation and the evolution of pathogen virulence. An important lesson for the study and management of future disease outbreaks is not only that portable sequencing platforms should be deployed as rapidly as possible, but that each sequence obtained should be linked to as much relevant metadata as is ethically and technically possible.

Despite the insights provided by the analysis of EBOV genome data, it is also clear that major questions remain. For future outbreaks it will be important to resolve exact chains of transmission (that is, who infected whom), as this provides vital information on the patterns and mechanisms of virus spread within single communities and hospitals, which will help target interventions. Sequence data from the 2013–2016 EVD epidemic indicated that these chains were difficult to infer using the population consensus sequences from individual hosts, although in several cases they were shown to be in agreement with epidemiological studies<sup>2,3,25</sup>. Hence, although EBOV evolves rapidly, mutations are not necessarily fixed at the scale of individual transmission events, which limits phylogenetic resolution. One solution is to examine the transmission patterns of intra-host single nucleotide variants (iSNVs) (ref. 3). If multiple iSNVs are routinely transmitted between individuals (that is, that EBOV is not subject to a severe population bottleneck at inter-host transmission) then tracking the inheritance patterns of these variants can provide information on how transmission patterns exist between individual hosts, as previously shown for influenza virus<sup>95,96</sup>. Importantly, it has been shown that substantial intra-host variation can be observed for EBOV, with 2–5 iSNVs per infected patient being typical when using a minor allele frequency cutoff of 5%<sup>2–4,26</sup>. As the cutoff is lowered, the numbers of observed iSNVs increase sharply<sup>2,3</sup>.

Genomic studies undertaken in West Africa towards the end of the 2013–2016 EVD outbreak illustrated how iSNV data can help to resolve EBOV transmission pathways. For example, Arias and colleagues showed how the analysis of iSNVs from EBOV patients in Sierra Leone could provide strong support for sexual transmission from EVD survivors<sup>26</sup>. Determining the number of iSNVs that transmit between hosts can also provide key information on the severity of the transmission bottleneck<sup>3</sup>, itself critical for understanding the ability of natural selection to shape patterns of genetic diversity.

Finally, while large-scale EBOV sequence studies have now been undertaken in human populations, there is an evident and critical need to determine the ecology and evolution of EBOV in its animal reservoir(s). While most current data points to bats being the ultimate reservoir host<sup>34,35</sup>, long-term studies of EBOV in bats have yet to be performed and it is likely that other host species exist, which may have a major bearing on epidemiological dynamics. To truly understand the ecology and evolution of EBOV, as well as its mechanisms of pathogenicity, will require information on the virus in all its host–virus interactions, and not just those associated with EVD outbreaks in humans.

Received 15 January; accepted 23 August 2016.

1. Baize, S. *et al.* Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.* **371**, 1418–1425 (2014).  
**The first paper to describe the emergence of EBOV Makona in Guinea in December 2013, providing the sequences of three full-length viral genomes.**
2. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).  
**Obtained the first large-scale, near-real-time EBOV genomic data from 78 patients in Sierra Leone, which provided critical insights into virus spread during the early stages of the epidemic.**
3. Park, D. J. *et al.* Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell* **161**, 1516–1526 (2015).
4. Andersen, K. G. *et al.* Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell* **162**, 738–750 (2015).
5. Kuhn, J. H. *et al.* Nomenclature- and database-compatible names for the two Ebola virus variants that emerged in Guinea and the Democratic Republic of the Congo in 2014. *Viruses* **6**, 4760–4799 (2014).
6. WHO. Statement on the 1st meeting of the IHR Emergency Committee on the 2014 Ebola outbreak in West Africa. *WHO Media Centre* <http://www.who.int/mediacentre/news/statements/2014/ebola-20140808/en/> (2014).
7. Ebola Situation Report, W. H. O. *Ebola virus disease outbreak* <http://apps.who.int/ebola/current-situation/ebola-situation-report-30-march-2016> (2016).
8. Nichol, S. T. *et al.* Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness. *Science* **262**, 914–917 (1993).
9. Holmes, E. C. *et al.* The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *J. Infect. Dis.* **171**, 45–53 (1995).
10. Tsui, S. K., Chim, S. S. & Lo, Y. M. Coronavirus genomic-sequence variations and the epidemiology of the severe acute respiratory syndrome. *N. Engl. J. Med.* **349**, 187–188 (2003).
11. Ghedin, E. *et al.* Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* **437**, 1162–1166 (2005).
12. Garten, R. J. *et al.* Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. *Science* **325**, 197–201 (2009).
13. Chiu, C. Y. Viral pathogen discovery. *Curr. Opin. Microbiol.* **16**, 468–478 (2013).
14. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
15. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
16. Mardis, E. R. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6**, 287–303 (2013).
17. Hoenen, T. *et al.* Virology. Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science* **348**, 117–119 (2015).
18. Bell, A. *et al.* Genome sequence analysis of Ebola virus in clinical samples from three British healthcare workers, August 2014 to March 2015. *Euro Surveill.* **20**, 21131 (2015).
19. Kugelman, J. R. *et al.* Monitoring of Ebola virus Makona evolution through establishment of advanced genomic capability in Liberia. *Emerg. Infect. Dis.* **21**, 1135–1143 (2015).
20. Simon-Loriere, E. *et al.* Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature* **524**, 102–104 (2015).
21. Carroll, M. W. *et al.* Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature* **524**, 97–101 (2015).  
**The first study to obtain a large EBOV sequence data set collected over multiple outbreaks, providing insights into the phylogenetic relationship between EBOV outbreak variants as well as other filoviruses.**
22. Tong, Y. G. *et al.* Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature* **524**, 93–96 (2015).
23. Smits, S. L. *et al.* Genotypic anomaly in Ebola virus strains circulating in Magazine Wharf area, Freetown, Sierra Leone, 2015. *Euro Surveill.* **20**, 30035 (2015).
24. Ladner, J. T. *et al.* Evolution and spread of Ebola virus in Liberia, 2014–2015. *Cell Host Microbe* **18**, 659–669 (2015).
25. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).  
**Implemented real-time sequencing in Guinea using the Oxford Nanopore MinION technology, which enabled EBOV genomic sequence data to be obtained within hours or days of cases being detected.**
26. Arias, A. *et al.* Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* **2**, vew016 (2016).  
**The largest single study of EBOV genomics to date. Implemented real-time sequencing in Sierra Leone using the Thermo Fisher Ion Torrent platform, which enabled EBOV genomic sequence data to be obtained within days of cases being detected.**
27. Matrangola, C. B. *et al.* Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* **15**, 519 (2014).
28. Woolhouse, M. E., Rambaut, A. & Kellam, P. Lessons from Ebola: improving infectious disease surveillance to inform outbreak management. *Sci. Transl. Med.* **7**, 307rv5 (2015).
29. Neher, R. A. & Bedford, T. Nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics* **31**, 3546–3548 (2015).
30. Bedford, T. & Neher, R. A. NextStrain—real-time analysis of Ebola virus evolution. *NextStrain* <http://ebola.nextstrain.org/> (2015).
31. Lewis, J. A. *et al.* Phylogenetic relationships of dengue-2 viruses. *Virology* **197**, 216–224 (1993).
32. Ray, S. C., Arthur, R. R., Carella, A., Bukh, J. & Thomas, D. L. Genetic epidemiology of hepatitis C virus throughout Egypt. *J. Infect. Dis.* **182**, 698–707 (2000).
33. WHO. WHO experts consultation on Ebola Reston pathogenicity in humans. *Emergencies Preparedness, Response* [http://www.who.int/csr/resources/publications/WHO\\_HSE\\_EPR\\_2009\\_2/en/](http://www.who.int/csr/resources/publications/WHO_HSE_EPR_2009_2/en/) (2015).
34. Leroy, E. M. *et al.* Fruit bats as reservoirs of Ebola virus. *Nature* **438**, 575–576 (2005).  
**Obtained sequence information from three species of fruit bats collected during the 2001–2003 EVD outbreak in Gabon to provide the first convincing evidence implicating fruit bats as a/the likely EBOV reservoir.**
35. Leendertz, S. A., Gogarten, J. F., Düx, A., Calvignac-Spencer, S. & Leendertz, F. H. Assessing the evidence supporting fruit bats as the primary reservoirs for Ebola viruses. *EcoHealth* **13**, 18–25 (2016).
36. Taylor, D. J., Leach, R. W. & Bruenn, J. Filoviruses are ancient and integrated into mammalian genomes. *BMC Evol. Biol.* **10**, 193 (2010).
37. Taylor, D. J., Ballinger, M. J., Zhan, J. J., Hanzly, L. E. & Bruenn, J. A. Evidence that ebolaviruses and cuevaviruses have been diverging from marburgviruses since the Miocene. *PeerJ* **2**, e556 (2014).
38. Report of an International Commission. Ebola haemorrhagic fever in Zaire, 1976. *Bull. World Health Organ.* **56**, 271–293 (1978).
39. WHO Ebola Response Team. Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371**, 1481–1495 (2014).
40. Kucharski, A. J. & Edmunds, W. J. Case fatality rate for Ebola virus disease in West Africa. *Lancet* **384**, 1260 (2014).
41. CDC. Outbreaks chronology: Ebola virus disease. *Outbreaks* <http://www.cdc.gov/vhf/ebola/outbreaks/history/chronology.html> (2015).
42. Marzi, A. *et al.* Delayed disease progression in Cynomolgus macaques infected with Ebola virus Makona strain. *Emerg. Infect. Dis.* **21**, 1777–1783 (2015).
43. Schieffelin, J. S. *et al.* Clinical illness and outcomes in patients with Ebola in Sierra Leone. *N. Engl. J. Med.* **371**, 2092–2100 (2014).
44. Volz, E. & Pond, S. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Curr.* **6**, ecurrents.outbreaks.6f7025f1271821d4c815385b08f5f80e (2014).
45. Alizon, S., Lion, S., Murali, C. L. & Abbate, J. L. Quantifying the epidemic spread of Ebola virus (EBOV) in Sierra Leone using phylodynamics. *Virulence* **5**, 825–827 (2014).
46. Althaus, C. L. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLoS Curr.* **6**, ecurrents.outbreaks.91afb5e0f279e7f29e7056095255b288 (2014).
47. Rosello, A. *et al.* Ebola virus disease in the Democratic Republic of the Congo, 1976–2014. *eLife* **4**, e09015 (2015).
48. Dudas, G. & Rambaut, A. Phylogenetic analysis of Guinea 2014 EBOV ebolavirus outbreak. *PLoS Curr.* **6**, ecurrents.outbreaks.84eef5ce43ec9dc0bf0670f7b8b417d (2014).
49. Leroy, E. M. *et al.* Multiple Ebola virus transmission events and rapid decline of central African wildlife. *Science* **303**, 387–390 (2004).
50. Wittmann, T. J. *et al.* Isolates of Zaire ebolavirus from wild apes reveal genetic lineage and recombinants. *Proc. Natl Acad. Sci. USA* **104**, 17123–17127 (2007).
51. Walsh, P. D., Biek, R. & Real, L. A. Wave-like spread of Ebola Zaire. *PLoS Biol.* **3**, e371 (2005).
52. Carroll, S. A. *et al.* Molecular evolution of viruses of the family *Filoviridae* based on 97 whole-genome sequences. *J. Virol.* **87**, 2608–2616 (2013).
53. Biek, R., Walsh, P. D., Leroy, E. M. & Real, L. A. Recent common ancestry of Ebola Zaire virus found in a bat reservoir. *PLoS Pathog.* **2**, e90 (2006).  
**Using previously available sequence data, this paper describes a strong link between EBOV found in bats and EBOV found in humans during outbreaks, suggesting common ancestry within the last three decades.**
54. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
55. Sanchez, A., Trappier, S. G., Mahy, B. W., Peters, C. J. & Nichol, S. T. The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing. *Proc. Natl Acad. Sci. USA* **93**, 3602–3607 (1996).
56. Georges-Courbot, M. C. *et al.* Isolation and phylogenetic characterization of Ebola viruses causing different outbreaks in Gabon. *Emerg. Infect. Dis.* **3**, 59–62 (1997).
57. Georges, A. J. *et al.* Ebola hemorrhagic fever outbreaks in Gabon, 1994–1997: epidemiologic and health control issues. *J. Infect. Dis.* **179** (Suppl. 1), S65–S75 (1999).
58. Rodriguez, L. L. *et al.* Persistence and genetic stability of Ebola virus during the outbreak in Kikwit, Democratic Republic of the Congo, 1995. *J. Infect. Dis.* **179** (Suppl. 1), S170–S176 (1999).
59. Grard, G. *et al.* Emergence of divergent Zaire ebola virus strains in Democratic Republic of the Congo in 2007 and 2008. *J. Infect. Dis.* **204** (Suppl. 3), S776–S784 (2011).
60. Łuksza, M., Bedford, T. & Lässig, M. Epidemiological and evolutionary analysis of the 2014 Ebola virus outbreak. Preprint at: <http://arxiv.org/pdf/1411.1722.pdf> (2014).
61. Holmes, E. C. *The Evolution and Emergence of RNA Viruses*. (Oxford Univ. Press, 2009).



62. Osterholm, M. T. *et al.* Transmission of Ebola viruses: what we know and what we do not know. *MBio* **6**, e00137 (2015).
63. Ponce De Leon-Rosales, S., Arredondo-Hernandez, R., Macias, A. & Wenzel, R. P. Ebola, through air or not through air: that is the question. *Front. Public Health* **2**, 292 (2015).
64. Leroy, E. M., Labouba, I., Maganga, G. D. & Berthet, N. Ebola in West Africa: the outbreak able to change many things. *Clin. Microbiol. Infect.* **20**, 0597–0599 (2014).
65. Osterholm, M. T. What we're afraid to say about Ebola. *NY Times* (11 Sept. 2014).
66. Basch, C. H., Basch, C. E. & Redlener, I. Coverage of the Ebola virus disease epidemic in three widely circulated United States newspapers: implications for preparedness and prevention. *Health Promot. Perspect.* **4**, 247–251 (2014).
67. Callaway, E. M. Ebola's fast evolution questioned. *NATNEWS* doi:10.1038/nature.2015.17200 (2015).
68. Vogel, G. A reassuring snapshot of Ebola. *Science* **347**, 1407 (2015).
69. Saphire, E. O. New advances in the effort against Ebola. *Cell Host Microbe* **17**, 545–547 (2015).
70. Liu, S. Q., Rayner, S. & Zhang, B. How Ebola has been evolving in West Africa. *Trends Microbiol.* **23**, 387–388 (2015).
71. Ho, S. Y., Shapiro, B., Phillips, M. J., Cooper, A. & Drummond, A. J. Evidence for time dependency of molecular rate estimates. *Syst. Biol.* **56**, 515–522 (2007).
72. Duchêne, S., Holmes, E. C. & Ho, S. Y. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. *Proc. R. Soc.* **281**, 20140732 (2014).
73. Moya, A., Holmes, E. C. & González-Candelas, F. The population genetics and evolutionary epidemiology of RNA viruses. *Nat. Rev. Microbiol.* **2**, 279–288 (2004).
74. Maganga, G. D. *et al.* Ebola virus disease in the Democratic Republic of Congo. *N. Engl. J. Med.* **371**, 2083–2091 (2014).
75. Lam, T. T., Zhu, H., Chong, Y. L., Holmes, E. C. & Guan, Y. Puzzling origins of the Ebola outbreak in the Democratic Republic of the Congo, 2014. *J. Virol.* **89**, 10130–10132 (2015).
76. Blackley, D. J. *et al.* Reduced evolutionary rate in reemerged Ebola virus transmission chains. *Sci. Adv.* **2**, e1600378 (2016).
- Using EBOV sequencing from seven EVD cases collected during a 'flare-up' in Liberia, this study investigates how these cases were linked to individuals who had survived EVD and carried the virus as a persistent asymptomatic infection.**
77. Mate, S. E. *et al.* Molecular evidence of sexual transmission of Ebola virus. *N. Engl. J. Med.* **373**, 2448–2454 (2015).
- Using a combination of epidemiological and sequence-based investigations, this study convincingly shows that EBOV is capable of sexual transmission many months after an EVD survivor was discharged from the hospital.**
78. Malpica, J. M. *et al.* The rate and character of spontaneous mutation in an RNA virus. *Genetics* **162**, 1505–1511 (2002).
79. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84**, 9733–9748 (2010).
80. Domingo-Calap, P. & Sanjuán, R. Experimental evolution of RNA versus DNA viruses. *Evolution* **65**, 2987–2994 (2011).
81. Kuhn, J. H. Filoviruses. A compendium of 40 years of epidemiological, clinical, and laboratory studies. *Arch. Virol. Suppl.* **20**, 13–360 (2008).
82. Côté, M. *et al.* Small molecule inhibitors reveal Niemann-Pick C1 is essential for Ebola virus infection. *Nature* **477**, 344–348 (2011).
83. Carette, J. E. *et al.* Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477**, 340–343 (2011).
84. Horimoto, T. & Kawaoka, Y. Influenza: lessons from past pandemics, warnings from current incidents. *Nat. Rev. Microbiol.* **3**, 591–600 (2005).
85. Deen, G. F. *et al.* Ebola RNA persistence in semen of Ebola virus disease survivors—preliminary report. *N. Engl. J. Med.* (2015). 10.1056/NEJMoa1511410
86. Christie, A. *et al.* Possible sexual transmission of Ebola virus—Liberia, 2015. *MMWR Morb. Mortal. Wkly. Rep.* **64**, 479–481 (2015).
87. Fischer, R. J., Judson, S., Miazgowiec, K., Bushmaker, T. & Munster, V. J. Ebola virus persistence in semen *ex vivo*. *Emerg. Infect. Dis.* **22**, 289–291 (2016).
88. Thorson, A., Formenty, P., Lofthouse, C. & Broutet, N. Systematic review of the literature on viral persistence and sexual transmission from recovered Ebola survivors: evidence and recommendations. *BMJ Open* **6**, e008859 (2016).
89. Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. Bayesian phylogeography finds its roots. *PLOS Comput. Biol.* **5**, e1000520 (2009).
90. Stadler, T. *et al.* Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol.* **29**, 347–357 (2012).
91. Stadler, T., Kühnert, D., Rasmussen, D. A. & du Plessis, L. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS Curr.* **6**, ecurrents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f (2014).
- Using sequence data from 72 EVD cases from Sierra Leone, this study utilized phylodynamic analyses to estimate several epidemiological parameters, obtaining an  $R_0$  of around 2.18.**
92. Stadler, T. & Bonhoeffer, S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. R. Soc. Lond. B* **368**, 20120198 (2013).
93. Hoenen, T. *et al.* Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerg. Infect. Dis.* **22**, 331–334 (2016).
94. Yozwiak, N. L. *et al.* Roots, not parachutes: research collaborations combat outbreaks. *Cell* **166**, 5–8 (2016).
95. Poon, L. L. *et al.* Quantifying influenza virus diversity and transmission in humans. *Nat. Genet.* **48**, 195–200 (2016).
96. Stack, J. C., Murcia, P. R., Grenfell, B. T., Wood, J. L. & Holmes, E. C. Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proc. R. Soc.* **280**, 20122173 (2013).
97. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
98. Hoenen, T. *et al.* Erratum: Mutation rate and genotype variation of Ebola virus from Mali case sequences. *Science* **348**, aac5674 (2015).
99. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21**, 456–463 (2005).
100. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).

**Acknowledgements** We thank all the African doctors, nurses, scientists, and outbreak responders who worked to control the 2013–2016 EVD epidemic, some of whom tragically died in the process. We also thank the EBOV genome sequence data producers for making their data publicly available, S. Schaffner for suggestions and reading of the manuscript, and L. M. Carvalho for donating evolutionary rate data. E.C.H. is funded by an NHMRC Australia Fellowship (AF30). G.D. is supported by EU (FP7/2007–2013) Grant Agreement no. 278433-PREDEMICS and the Mahan Postdoctoral Fellowship from the Computational Biology Program at Fred Hutchinson Cancer Research Center. A.R. is supported by EU (FP7/2007–2013) Grant Agreement no. 278433-PREDEMICS, H2020 Grant Agreement no. 643476-COMPARE, and a Wellcome Trust Strategic Award (VIZIONS; 093724). K.G.A. is a PEW Biomedical Scholar, and his work is supported by an NIH National Center for Advancing Translational Studies Clinical and Translational Science Award UL1TR001114, and NIAID contract HHSN272201400048C.

**Author Contributions** All authors were involved in data analysis and writing of the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.C.H. ([edward.holmes@sydney.edu.au](mailto:edward.holmes@sydney.edu.au)) or K.G.A. ([kristian@andersen-lab.com](mailto:kristian@andersen-lab.com)).

**Reviewer Information** *Nature* thanks C. Drosten, P. Lemey, G. Palacios and T. Sadler for their contribution to the peer review of this work.

# The Simons Genome Diversity Project: 300 genomes from 142 diverse populations

A list of authors and affiliations appears at the end of the paper.

**Here we report the Simons Genome Diversity Project data set: high quality genomes from 300 individuals from 142 diverse populations. These genomes include at least 5.8 million base pairs that are not present in the human reference genome. Our analysis reveals key features of the landscape of human genome variation, including that the rate of accumulation of mutations has accelerated by about 5% in non-Africans compared to Africans since divergence. We show that the ancestors of some pairs of present-day human populations were substantially separated by 100,000 years ago, well before the archaeologically attested onset of behavioural modernity. We also demonstrate that indigenous Australians, New Guineans and Andamanese do not derive substantial ancestry from an early dispersal of modern humans; instead, their modern human ancestry is consistent with coming from the same source as that of other non-Africans.**

To obtain a complete picture of human diversity, it is necessary to sequence the genomes of many individuals from diverse locations. To date, the largest whole-genome sequencing survey, the 1000 Genomes Project, analysed 26 populations of European, East Asian, South Asian, American, and sub-Saharan African ancestry<sup>1</sup>. However, this and most other sequencing studies have focused on demographically large populations. Such studies tend to ignore smaller populations that are also important for understanding human diversity. In addition, many of these studies have sequenced genomes to only 4–6-fold coverage. Here, we report the Simons Genome Diversity Project (SGDP): deep genome sequences of 300 individuals from 142 populations chosen to span much of human genetic, linguistic, and cultural variation (Supplementary Data Table 1).

## Data set and catalogue of novel variants

We sequenced the samples to an average coverage of 43-fold (range 34–83-fold) at Illumina Ltd; almost all samples (278) were prepared using the same PCR-free library preparation ([https://support.illumina.com/content/dam/illumina-marketing/documents/services/FastTrackServices\\_Methods\\_Tech\\_Note.pdf](https://support.illumina.com/content/dam/illumina-marketing/documents/services/FastTrackServices_Methods_Tech_Note.pdf)). We aligned reads to the human reference genome hs37d5/hg19 using BWA-MEM (BWA-0.7.12)<sup>2</sup> (Supplementary Information section 1). We genotyped each sample separately using the Genome Analysis Toolkit (GATK)<sup>3</sup>, with a modification to eliminate bias towards genotypes matching the reference (Supplementary Information section 1). We developed a filtering procedure that generates a sample-specific mask. At ‘filter level 1’ which we recommend for most analyses, we retain an average of 2.13 Gb of sequence per sample and identify 34.4 million single nucleotide polymorphisms (SNPs) and 2.1 million insertion/deletion polymorphisms (indels) (Supplementary Information section 2). We have made the GATK-processed data available in a file small enough to download by FTP, along with software to analyse these data (Supplementary Information section 3). The SGDP data set highlights the incompleteness of current catalogues of human variation, with the fraction of heterozygous positions not discovered by the 1000 Genomes Project being 11% in the KhoeSan and 5% in New Guineans and Australians (Extended Data Fig. 1; Supplementary Data Table 1). We used FermiKit<sup>4</sup> to map short reads against each other, store the assemblies in a compressed form that retains all the information

required for polymorphism discovery and analysis, and identified SNPs by comparing against the human reference. We find that FermiKit has comparable sensitivity and specificity to GATK for SNP discovery and genotyping, and is more accurate for indels (Supplementary Information section 4). FermiKit also identified 5.8 Mb of contigs that are present in the SGDP but absent in the human reference genome presumably because they are deleted there; these contigs which we have made publicly available can be used as ‘decoys’ to improve read mapping (Supplementary Information section 5). Finally, we called copy number variants<sup>5</sup> and used lobSTR<sup>6,7</sup> to genotype 1.6 million short tandem repeats (STRs) (Supplementary Information section 6). The high quality of the STR genotypes ( $r^2 = 0.92$  to capillary sequencing calls) is evident from their accurate reconstruction of population relationships, even for difficult-to-genotype mononucleotide repeats (Extended Data Fig. 2).

## The structure of human genetic diversity

To obtain an overview of population relationships, we carried out ADMIXTURE<sup>8</sup> (Extended Data Fig. 3) and principal component analysis<sup>9</sup> (Extended Data Fig. 4a). We also built neighbour-joining trees based on pairwise divergence per nucleotide (Fig. 1a) and  $F_{ST}$  (Extended Data Fig. 4b) whose topologies are consistent with previous findings that the deepest splits among human populations are among Africans. We computed heterozygosity—the proportion of diallelic genotypes per base pair—and recapitulate previous findings that the highest genetic diversity is found in sub-Saharan Africa and that there is a much lower ratio of X-to-autosome diversity in non-Africans than in Africans (Fig. 1b)<sup>10</sup>. A surprise is that African ‘pygmy’ hunter-gatherers have reduced X-to-autosome diversity ratios relative to all other sub-Saharan Africans. This pattern is just as strong even after we remove the third of chromosome X known to be subject to the strongest natural selection, suggesting that the finding is driven by demographic history rather than by natural selection (Supplementary Information section 7). It has been suggested that the reduced X-to-autosome heterozygosity ratio in non-Africans is due to ongoing male-driven admixture<sup>10,11</sup>. Male non-pygmy admixture into pygmies is well-documented<sup>12,13</sup>, so this process could explain these findings.

Comparisons of ancient to present-day human genomes have shown that all non-Africans today possess Neanderthal ancestry<sup>14</sup> with

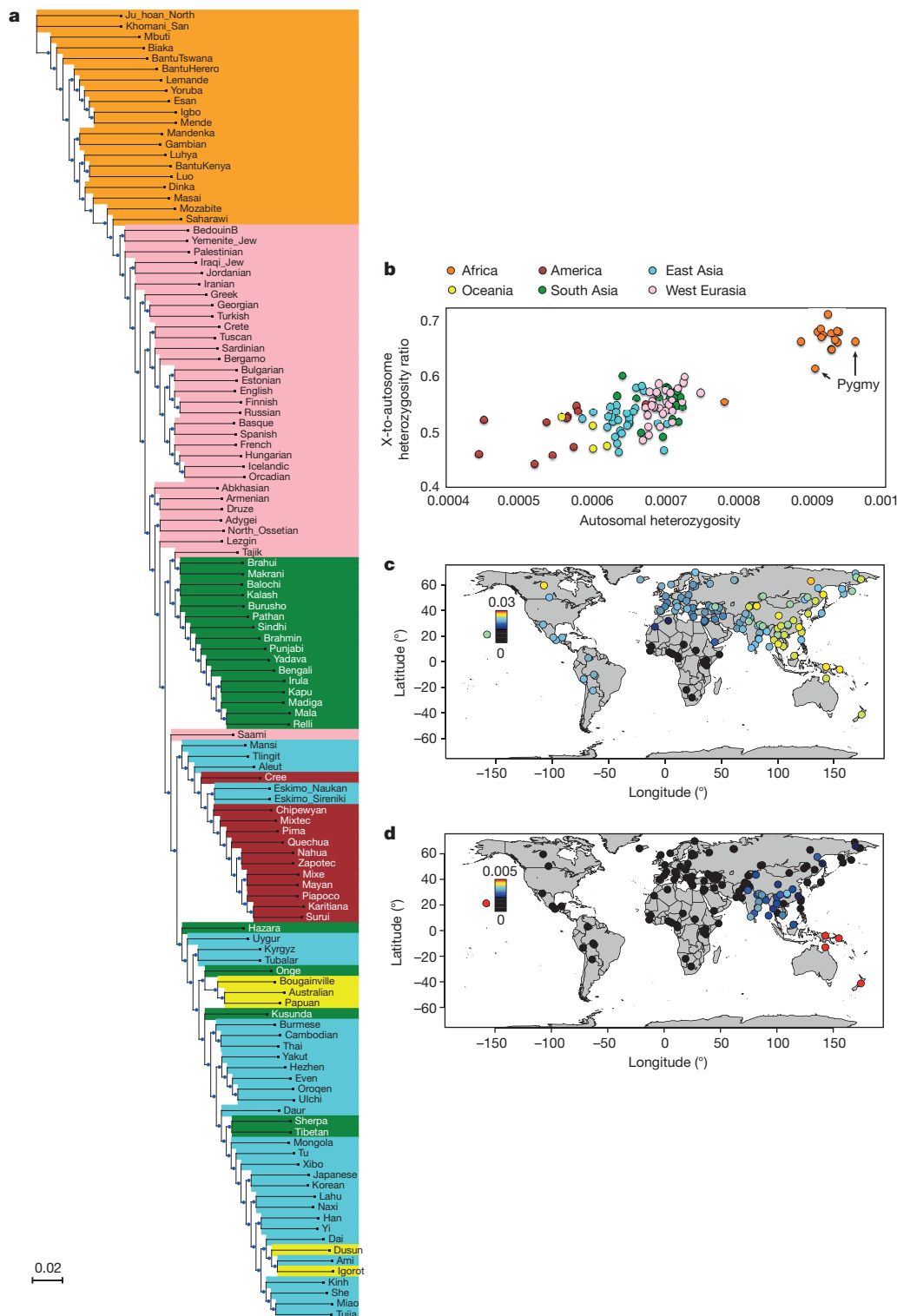


more in eastern non-Africans<sup>15,16</sup>, and that Australo-Melanesians, and to a lesser extent other eastern non-Africans, possess Denisovan ancestry<sup>17–19</sup>. However, these studies only analysed genomes from a handful of populations. We computed statistics informative about Neanderthal and Denisovan ancestry and provide a fine-scale view of these ancestry distributions worldwide (Fig. 1c, d; Supplementary Data Table 1; Supplementary Information section 8). We do not detect any population with a higher proportion of Neanderthal ancestry than is present in East Asians. However, we do find suggestive evidence of an excess of Denisovan ancestry in some South Asians compared to other Eurasians. This signal may not have been detected before because

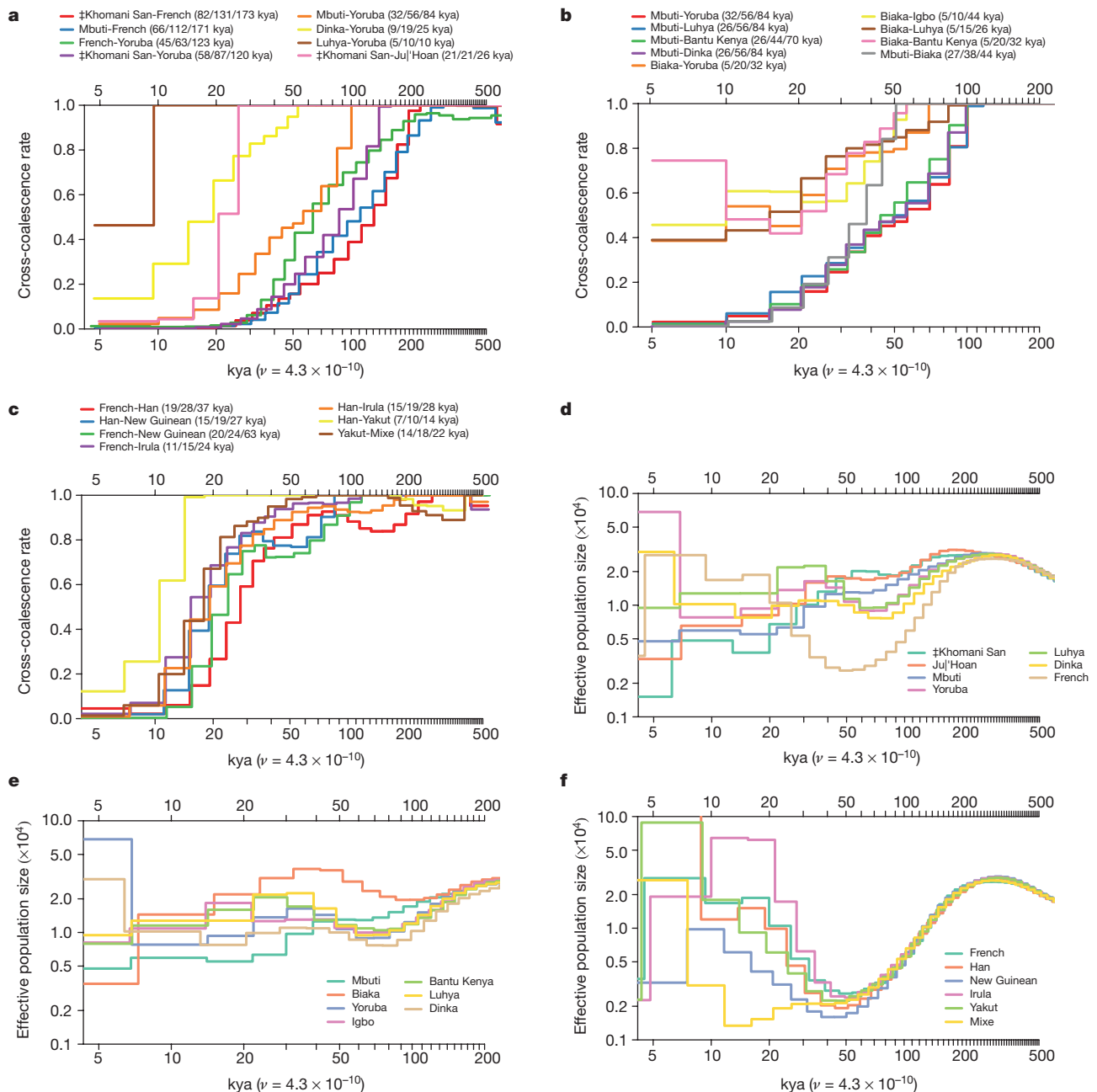
earlier surveys of archaic introgression largely excluded South Asians (Fig. 1d; Supplementary Data Table 1).

### The time course of human population separation

We studied demographic history by leveraging the fact that variation across the genome in divergent sites per base pair can be used to reconstruct population size changes and separations. We used the pairwise sequential Markovian coalescent (PSMC)<sup>20</sup> to reconstruct population size changes, and the multiple sequentially Markovian coalescent (MSMC)<sup>21</sup> to study the time course of population separations. We infer that the population ancestral to all present day humans began to develop



**Figure 1 | Genetic variation in the SGDP. a**, Neighbour-joining tree of relationships based on pairwise divergence. **b**, Plot of autosomal heterozygosity against the X-to-autosome heterozygosity ratio, showing the reduction in this ratio in non-Africans and pygmies. **c**, Estimate of Neanderthal ancestry with a heat map scale of 0–3%. **d**, Estimate of Denisovan ancestry with a heat map scale of 0–0.5% to bring out subtle differences in mainland Eurasia (Oceania groups with as much as 5% Denisovan ancestry are saturated in bright red).



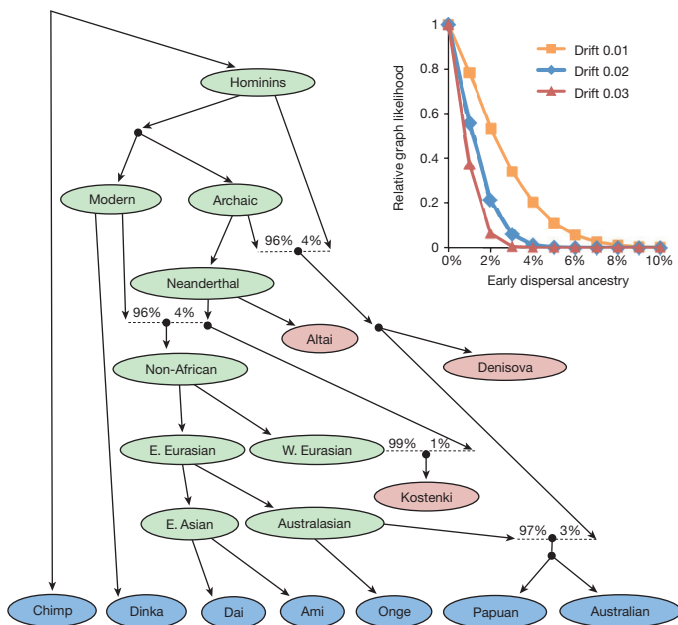
**Figure 2 | Cross-coalescence rates and effective population sizes for selected population pairs. a–c,** Cross-coalescence rates as a function of time in thousands of years ago (kya) estimated using MSMC, with four haplotypes per pair. In each subfigure legend, we give the point estimate of the date at which 25%, 50% and 75% of lineages in the pair of populations have coalesced into a common ancestral population. We generated these plots using data phased with the 1000 Genomes reference panel (method

PS1 described in Supplementary Information section 9), but only show pairs of populations for which the cross-coalescence rates are relatively insensitive to the phasing approach. **a,** Selected African cross-coalescence rates. **b,** Central African rainforest hunter-gatherer cross-coalescence rates. **c,** Ancient non-African cross-coalescence rates. **d–f,** Effective population sizes inferred using PSMC, using one diploid genome per population, for the same populations that we used in **a–c**.

substructure at least 200 thousand years ago (kya), which is most apparent when comparing the ancestors of some present-day African hunter-gatherers (southern African KhoeSan and central African Mbuti pygmies) to other populations (Fig. 2a). However, it is also clear that this substructure developed slowly, as all pairs of present-day populations including African hunter-gatherers share a substantial subset of their ancestors as recently as a hundred thousand years ago<sup>22–25</sup>. Quoting the time at which MSMC infers that more than 50% (25–75%) of lineages for a pair of populations are descended from the same ancestral population, we estimate that non-Africans separated substantially from KhoeSan 131 (82–173) kya and almost as anciently from the Mbuti around 112 (66–171) kya. Within Africa (Fig. 2a, b), we infer that

the Yoruba separated substantially from the KhoeSan 87 (58–120) kya; from the Mbuti 56 (32–84) kya; and from the Dinka 19 (9–25) kya. We estimate a relatively rapid 21 (21–26) kya separation of northern and southern KhoeSan<sup>23,26</sup> potentially reflecting isolation since the last glacial maximum; and 38 (27–44) kya separation between western (Biaka) and eastern (Mbuti) pygmies, confirming very old substructure between these two central African hunter-gatherer groups<sup>27</sup>. Outside Africa, the most ancient structure dates to around 50 kya (Fig. 2c) during or shortly after the deepest part of the shared non-African bottleneck 40–60 kya, consistent with the archaeological evidence of the dispersal of modern humans into Eurasia during this period. We are not confident about the estimates of the date of





**Figure 3 | Present-day populations have negligible ancestry from an early dispersal of modern humans out of Africa.** Best-fitting admixture graph model of relationships among Australians, New Guineans, Andamanese and other diverse populations. Present-day populations are shown in blue, ancient samples in red, and select inferred ancestral nodes in green. Dotted lines indicate admixture events, all of which involve archaic humans. All  $f$ -statistic relationships are accurately fit to within 2.1 standard errors. Inset, results of adding putative early dispersal admixture to the graph model for different assumptions about when the early lineage split off. We specify the split time in terms of the genetic drift above the 'Non-African' node, with 0.01 units of drift representing on the order of ten thousand years. The (approximate) model likelihood is maximized with zero early dispersal ancestry, and no more than a few per cent is consistent with the data.

separation of Australians, New Guineans and Andamanese from other populations because we find that these inferences change depending on the computational method we use for phasing, probably due to these populations not being represented in the 1000 Genomes haploid genome reference panel (Supplementary Information section 9). We caution that the date estimates also do not take into account uncertainty about the true value of the human mutation rate, which could plausibly be 30% higher or lower than the point estimate we use<sup>28</sup>.

### Early modern human dispersals contributed little to non-Africans

There is intense debate about whether present-day Australians, New Guineans and Asian 'Negrito' populations are descended from the same source population as mainland Eurasians, or whether they also derive some ancestry from an early, independent dispersal of modern humans into Asia<sup>29–31</sup>. To explore this scenario rigorously, we fit an admixture graph<sup>32</sup>—a phylogenetic tree incorporating mixture events—to the allele frequency correlations among Neanderthals, Denisovans, Upper Paleolithic Europeans, East Asians, New Guineans, Australians, and Andamanese. We obtain a good fit to the data if we include known Neanderthal and Denisovan introgression and model all modern human ancestry in New Guineans, Australians and Andamanese as part of an eastern clade together with mainland East Asians (Supplementary Information section 11; Fig. 3). Furthermore, when we manually introduce a deeply diverging modern human lineage contributing ancestry to Australians, New Guineans, and Andamanese (or when we repeat the analysis in a model without Andamanese), no position or proportion of the deep lineage improves the fit. If this putative source population branched off the main lineage leading to

non-Africans more than about 10–20 thousand years before the separation of European and East Asian ancestors, we obtain an upper bound of a few per cent for the possible contribution to Australians and New Guineans (Fig. 3 inset; Supplementary Information section 11). These results are at odds with an inference of substantial early dispersal ancestry in a previous analysis of an Australian genome<sup>31</sup>, however, that study used a less complete model that, notably, did not include the known Denisovan admixture into Australo-Melanesians<sup>17</sup>. The findings for Australians are also unlikely to be due to some unusual feature of the individuals we sequenced, as when we compared three different groups of Australian samples for which there is published genome-wide data, we found them all to be consistent with descending from a common homogeneous population since separation from New Guineans (Supplementary Information section 10). These results are not in conflict with skeletal and archaeological evidence of an early modern human presence outside of Africa<sup>29,33</sup>, as early migrations could have occurred but not contributed substantially to present-day populations. The possibility of populations that once flourished but did not contribute substantially to living groups is especially plausible now that ancient DNA from the ~45 kya Ust'-Ishim<sup>28</sup> and the ~40 kya Oase 1 individuals<sup>34</sup> has documented their existence.

### Accelerated mutation accumulation in non-Africans

The SGDP data provide an opportunity to compare the rates at which mutations have accumulated across populations. We restricted our analyses to samples for which our genotypes are likely to be most reliable (this included restricting to samples which were all processed in the same way), and we used the highest level of filtering ('level 9') (Supplementary Information section 7). We pooled samples by region to increase power, and for all pairs of regions, computed the expected number of positions where, if we picked a random chromosome from both, region A would mismatch chimpanzee and region B would be identical to chimpanzee (or vice versa). If the rate of accumulation of mutation has been the same since the two populations diverged, these numbers are expected to be equal<sup>35</sup>. However, when we compute the ratio of mutations on one lineage or the other since separation, we find a subtle (average of 0.5%) but significant excess of mutations in non-Africans relative to sub-Saharan Africans ( $3.3 < |Z| < 9.4$  standard errors from zero; Extended Data Table 1). Because any difference must reflect events since non-African/African population divergence, which is a less than a tenth of average genetic divergence (Fig. 2a), this implies a greater difference in mutation accumulation rates since population divergence (~5%). We were concerned that these results might be biased by the fact that the human genome reference sequence is more closely related to non-Africans than to Africans, or by higher levels of heterozygosity in Africans, as both of these issues could make detection of divergent sites in Africans more difficult. However, we replicated the findings after remapping to chimpanzee, which is equally distant to all present populations, and after restricting analyses to the X chromosome in males (as males only have a single X chromosome, this procedure avoids bias due to different error rates in detecting heterozygous genotypes in populations with different rates of heterozygosity) (Extended Data Fig. 5). These observations are most likely to be explained by acceleration in the rate of mutation accumulation in non-Africans, since the same signal appears in comparisons to sub-Saharan Africans related in different ways to non-Africans (Extended Data Table 1). It is known that the rate of CCT > CTT mutations differs across human populations. However, this particular mutation class was found to be enriched relative to Africans in Europeans but not in East Asians, and thus cannot explain our signal<sup>36</sup>. One of several possible explanations for these findings is a decrease in the generation interval in non-Africans compared to Africans since separation<sup>37</sup>.

### No species-wide sweeps in modern humans

Finally, we used the SGDP data set to address the hypothesis that the widespread appearance of modern human behaviour in the

archaeological record after ~50 kya was driven by one or a few changes in neurological genes that swept through the population shortly before this time<sup>38</sup>. We first applied the 3P-CLR method<sup>39</sup> to search for locations in the genome with low allele frequency differentiation between KhoeSan and other modern humans, combined with high differentiation between modern and archaic (Neanderthal and Denisovan) humans, as might be expected from a selective sweep in the ancestors of all modern humans (Supplementary Information section 12) (Extended Data Fig. 6). We found no strong outlier signals, although a caveat is that the scan has limited power and we could not apply it to filtered sections of the genome. We also applied the PSMC method<sup>20</sup> to estimate the average time since the most recent common ancestor (TMRCA) of individuals' two chromosomes in the genomic regions within the largest 3P-CLR peaks (38 peaks corresponding to the top 0.1%). In none of the regions did we infer that the great majority of all pairs of modern humans share a common ancestor <100 kya, as would be expected for a sweep just before ~50 kya years ago (Supplementary Data Table 2).

As a second approach to scanning for species-wide selective sweeps, we applied the PSMC to infer TMRCA for SGDP samples across the entire genome. This analysis found no regions where the great majority of pairs of human genomes are inferred to share a common ancestor <100 kya (the largest fraction seen anywhere in the genome is 68%; Extended Data Fig. 7).

Taken together, these results do not rule out the possibility that genetic changes contributed in a meaningful way to changes in human behaviour after 50 kya; for example, changing selection can produce shifts in the frequencies of pre-existing mutations to bring a population to a new and advantageous set-point for a phenotype as occurred in the case of height differences between northern and southern Europeans<sup>40</sup>. For polygenic selection, however, genetics is not a creative force, and instead responds to selection pressures imposed by novel environmental conditions or lifestyles. Thus, our results provide evidence against a model in which one or a few mutations were responsible for the rapid developments in human behaviour in the last 50,000 years. Instead, changes in lifestyles due to cultural innovation or exposure to new environments are likely to have been driving forces behind the rapid transformations in human behaviour in the last 50,000 years<sup>41,42</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 18 September 2015; accepted 23 June 2016.**

**Published online 21 September 2016.**

- Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
- McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
- Li, H. *FermiKit*: assembly-based variant calling for Illumina resequencing data. Preprint at <http://arxiv.org/abs/1504.06574> (2015).
- Sudmant, P. H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
- Gymrek, M. & Erlich, Y. Profiling short tandem repeats from short reads. *Methods Mol. Biol.* **1038**, 113–135 (2013).
- Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. *lobSTR*: A short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
- Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
- Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
- Keinan, A., Mullikin, J. C., Patterson, N. & Reich, D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat. Genet.* **41**, 66–70 (2009).
- Keinan, A. & Reich, D. Can a sex-biased human demography account for the reduced effective population size of chromosome X in non-Africans? *Mol. Biol. Evol.* **27**, 2312–2321 (2010).
- Verdu, P. *et al.* Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Mol. Biol. Evol.* **30**, 918–937 (2013).

- Joiris, D. V. The framework of central African hunter-gatherers and neighbouring societies. *African Study Monographs Suppl.* **28**, 57–79 (2003).
- Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
- Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
- Wall, J. D. *et al.* Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209 (2013).
- Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
- Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
- Skoglund, P. & Jakobsson, M. Archaic human ancestry in East Asia. *Proc. Natl Acad. Sci. USA* **108**, 18301–18306 (2011).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
- Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43**, 1031–1034 (2011).
- Schlebusch, C. M. *et al.* Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science* **338**, 374–379 (2012).
- Veeramah, K. R. *et al.* An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Mol. Biol. Evol.* **29**, 617–630 (2012).
- Labuda, D., Zietkiewicz, E. & Yotova, V. Archaic lineages in the history of modern humans. *Genetics* **156**, 799–808 (2000).
- Pickrell, J. K. *et al.* The genetic prehistory of southern Africa. *Nat. Commun.* **3**, 1143 (2012).
- Patin, E. *et al.* Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet.* **5**, e1000448 (2009).
- Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
- Groucutt, H. S. *et al.* Rethinking the dispersal of Homo sapiens out of Africa. *Evol. Anthropol.* **24**, 149–164 (2015).
- Reyes-Centeno, H., Hubbe, M., Hanihara, T., Stringer, C. & Harvati, K. Testing modern human out-of-Africa dispersal models and implications for modern human origins. *J. Hum. Evol.* **87**, 95–106 (2015).
- Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98 (2011).
- Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
- Liu, W. *et al.* The earliest unequivocally modern humans in southern China. *Nature* **526**, 696–699 (2015).
- Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
- Do, R. *et al.* No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet.* **47**, 126–131 (2015).
- Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl Acad. Sci. USA* **112**, 3439–3444 (2015).
- Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15**, 47–70 (2014).
- Klein, R. G. & Edgar, B. *The dawn of human culture.* (Wiley, 2002).
- Racimo, F. Testing for ancient selection using cross-population allele frequency differentiation. *Genetics* **202**, 733–750 (2015).
- Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44**, 1015–1019 (2012).
- McBrearty, S. & Brooks, A. S. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J. Hum. Evol.* **39**, 453–563 (2000).
- Renfrew, C. *Prehistory: the Making of the Human Mind.* (Modern Library, 2009).
- Alexander, D. H. & Lange, K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* **12**, 246 (2011).
- Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
- Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank the volunteers who donated samples. We thank H. Blanche, N. Boivin, H. Cann (deceased), E. Eichler, H. Greely, M. Petraglia, K. Prüfer, A. Rogers, M. Steinrücken, U. Stenzel and P. Sudmant for comments, critiques, discussions, or advice on assembling samples. We thank S. Fan for uploading 21 genomes to the European Genome-phenome archive. The sequencing was funded by the Simons Foundation (SFARI 280376) and the US National Science Foundation (BCS-1032255). I.M. was supported by a Long Term Fellowship grant LT001095/2014 from the Human Frontier Science program. P.S. was supported by the Wenner-Gren foundation and the Swedish Research Council (VR grant 2014-453). T.W. and M.G. were supported by an NIJ grant 2014-DN-BX-K089. Y.E. was supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and by NIJ grant 2014-DN-BX-K089. D.L. was supported by the Natural Sciences and Engineering



Research Council of Canada. T.K. was supported by ERC Starting Investigator grant FP7 - 261213. R.S. received support from Russian Foundation for Basic Research (#15-04-02543). S.D. received support from the Russian Foundation for Basic Research (#16-34-00599). R.K., E.K. and S.L. were supported by the Russian Foundation for Basic Research (11-04-00725-a). E.B. was supported by the Russian Foundation for Basic Research (16-06-00303). O.B. was supported by the Russian Scientific Fund (14-04-00827) and by the Russian Foundation for Basic Research (16-04-00890). D.M.B., H.S., E.M., R.V. and M.M. were supported by Institutional Research Funding from the Estonian Research Council IUT24-1 and by the European Regional Development Fund (European Union) through the Centre of Excellence in Genomics to Estonian Biocentre and University of Tartu. D.C. was supported by the Spanish MINECO grant CGL-44351-P. L.B.J. and W.S.W. were supported by NIH grant GM59290. S.A.T. was supported by NIH grants 5DP1ES022577 05, 1R01DK104339-01, and 1R01GM113657-01. C.T.-S. and Y.X. were supported by The Wellcome Trust grant 098051. C.M.B. was supported by NSF grants 0924726 and 1153911. K.T. was supported by CSIR Network Project grant (GENESIS: BSC0121). J.P.S. and Y.S.S. were supported in part by an NIH grant R01-GM094402, and a Packard Fellowship for Science and Engineering. G.R., J.K. and S.P. were funded by the Max Planck Society. N.P. and D.R. were supported by NIH grant GM100233 and D.R. is a Howard Hughes Medical Institute investigator.

**Author Contributions** S.M., Y.E., Y.S.S., S.P., J.K., N.P. and D.R. supervised the study. S.N., N.R., C.G., G.P., F.B., G.D., I.G.R., A.R.J., P.D., D.M.B., C.M.B., C.C., T.H., A.M.-E., O.L.P., E.B., O.B., S.K.-Y., H.S., D.T., L.Y., C.T.-S., Y.X., M.S.A., A.R.-L., C.B., A.D.R., C.J., E.B.S., E.M., J.P., R.V., B.M.H., U.H., R.W.M., A.S., G.S., J.T.S.W., R.K., E.K., S.L., G.A., D.C., M.H., T.K., W.K., C.A.W., D.L., M.B., L.B.J., S.A.T., W.S.W., M.M., S.D., R.S., L.S., K.T. and D.R. assembled samples. S.M., H.L., M.L., I.M., M.G., F.R., J.P.S., M.Z., N.C., A.T., P.S., I.L., S.S., Q.F., G.R., Y.S., N.P. and D.R. performed analyses. S.M., H.L., M.L., I.M., M.G., F.R., M.Z., N.P. and D.R. wrote the manuscript with help from all co-authors.

**Author Information** Raw data for 279 genomes for which the informed consent documentation is consistent with fully public data release are available through the EBI European Nucleotide Archive under accession numbers PRJEB9586 and ERP010710. For the remaining 21 genomes (designated by code 'Y' in the seventh column of Supplementary Data Table 1), data are deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001001959. Data for these 21 genomes can be obtained by submitting to the EGA Data Access Committee a signed letter containing the following text: "(a) I will not distribute the data outside my collaboration; (b) I will not post the data publicly; (c) I will make no attempt to connect the genetic data to personal identifiers for the samples; and (d) I will not use the data for any commercial purposes." Compact versions of the SGDP dataset and software for accessing it are available at ([http://genetics.med.harvard.edu/reichlab/Reich\\_Lab/Datasets.html](http://genetics.med.harvard.edu/reichlab/Reich_Lab/Datasets.html)). The short tandem repeat (STR) genotypes are available through dbVar under accession number nstd128 (<http://www.ncbi.nlm.nih.gov/dbvar>). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M. ([shop@genetics.med.harvard.edu](mailto:shop@genetics.med.harvard.edu)) or D.R. ([reich@genetics.med.harvard.edu](mailto:reich@genetics.med.harvard.edu)).

**Reviewer Information** *Nature* thanks P. Bellwood and S. Ramachandran and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Swapan Mallick<sup>1,2,3</sup>, Heng Li<sup>2\*</sup>, Mark Lipson<sup>1\*</sup>, Iain Mathieson<sup>1\*</sup>, Melissa Gymrek<sup>2,4,5,6</sup>, Fernando Racimo<sup>7</sup>, Mengyao Zhao<sup>1,2,3</sup>, Niru Chennagiri<sup>1,2,3</sup>, Susanne Nordenfelt<sup>1,2,3</sup>, Arti Tandon<sup>1,2</sup>, Pontus Skoglund<sup>1,2</sup>, Iosif Lazaridis<sup>1,2</sup>, Sriram Sankararaman<sup>1,2</sup>, Qiaomei Fu<sup>1,2,8</sup>, Nadin Rohland<sup>1,2</sup>, Gabriel Renaud<sup>9</sup>, Yaniv Erlich<sup>6,10,11</sup>, Thomas Willems<sup>6,12</sup>, Carla Gallo<sup>13</sup>, Jeffrey P. Spence<sup>14</sup>, Yun S. Song<sup>15,16,17</sup>, Giovanni Poletti<sup>13</sup>, Francois Balloux<sup>18</sup>, George van Driem<sup>19</sup>, Peter de Knijff<sup>20</sup>, Irene Gallego Romero<sup>21,22</sup>, Aashish R. Jha<sup>23</sup>, Doron M. Behar<sup>24</sup>, Claudio M. Bravi<sup>25</sup>, Cristian Capelli<sup>26</sup>, Tor Hervig<sup>27</sup>, Andres Moreno-Estrada<sup>28</sup>, Olga L. Posukh<sup>29,30</sup>, Elena Balanovska<sup>31</sup>, Oleg Balanovsky<sup>31,32,33</sup>, Sena Karachanak-Yankova<sup>34</sup>, Hovhannes Sahakyan<sup>24,35</sup>, Draga Toncheva<sup>34</sup>, Levon Yepiskoposyan<sup>35</sup>, Chris Tyler-Smith<sup>36</sup>, Yali Xue<sup>36</sup>, M. Syafiq Abdullah<sup>37</sup>, Andres Ruiz-Linares<sup>38</sup>, Cynthia M. Beall<sup>39</sup>, Anna Di Rienzo<sup>23</sup>, Choongwon Jeong<sup>23</sup>, Elena B. Starikovskaya<sup>40</sup>, Ene Metspalu<sup>24,41</sup>, Jüri Parik<sup>24</sup>, Richard Villems<sup>24,41,42</sup>, Brenna M. Henn<sup>43</sup>, Ugur Hodoglugli<sup>44</sup>, Robert Mahley<sup>45</sup>, Antti Sajantila<sup>46</sup>, George Stamatoyannopoulos<sup>47</sup>, Joseph T. S. Wee<sup>48</sup>, Rita Khusainova<sup>49,50</sup>, Elza Khusnutdinova<sup>49,50</sup>, Sergey Litvinov<sup>24,49,50</sup>, George Ayodo<sup>51</sup>, David Comas<sup>52</sup>, Michael F. Hammer<sup>53</sup>, Toomas Kivisild<sup>24,54</sup>, William Klitz<sup>6</sup>, Cheryl A. Winkler<sup>55</sup>, Damian Lubada<sup>56</sup>, Michael Bamshad<sup>57</sup>, Lynn B. Jorde<sup>58</sup>, Sarah A. Tishkoff<sup>59</sup>, W. Scott Watkins<sup>60</sup>, Mait Metspalu<sup>24</sup>, Stanislav Dryomov<sup>40,61</sup>, Rem Sukernik<sup>40,62</sup>, Lalji Singh<sup>63</sup>, Kumarasamy Thangaraj<sup>63</sup>, Svante Pääbo<sup>9</sup>, Janet Kelso<sup>9</sup>, Nick Patterson<sup>2</sup> & David Reich<sup>1,2,3</sup>

<sup>1</sup>Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

<sup>2</sup>Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. <sup>3</sup>Howard

Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA.

<sup>4</sup>Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA.

<sup>5</sup>Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, Massachusetts

02139, USA. <sup>6</sup>New York Genome Center, New York, New York 10013, USA. <sup>7</sup>Department

of Integrative Biology, University of California, Berkeley, California 94720-3140, USA. <sup>8</sup>Key

Laboratory of Vertebrate Evolution and Human Origins of Chinese Academy of Sciences,

IVPP, CAS, Beijing 100044, China. <sup>9</sup>Department of Evolutionary Genetics, Max Planck

Institute for Evolutionary Anthropology, D-04103 Leipzig, Germany. <sup>10</sup>Department of

Computer Science, Columbia University, New York, New York 10027, USA. <sup>11</sup>Center for

Computational Biology and Bioinformatics, Columbia University, New York, New York

10032, USA. <sup>12</sup>Computational and Systems Biology Program, Massachusetts Institute

of Technology, Cambridge, Massachusetts 02139, USA. <sup>13</sup>Laboratorios de Investigación

y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia,

Lima 15102, Perú. <sup>14</sup>Computational Biology Graduate Group, University of California,

Berkeley, California 94720, USA. <sup>15</sup>Computer Science Division, University of California,

Berkeley, California 94720, USA. <sup>16</sup>Department of Statistics, University of California,

Berkeley, California 94720, USA. <sup>17</sup>Department of Mathematics and Department of Biology,

University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>18</sup>Genetics Institute,

University College London, Gower Street, London WC1E 6BT, UK. <sup>19</sup>Institute of Linguistics,

University of Bern, Bern CH-3012, Switzerland. <sup>20</sup>Department of Human and Clinical

Genetics, Postzone S5-P, Leiden University Medical Center, 2333 ZA Leiden, Netherlands.

<sup>21</sup>School of Biological Sciences, Nanyang Technological University, 637551 Singapore. <sup>22</sup>Lee

Kong Chian School of Medicine, Nanyang Technological University, 636921 Singapore.

<sup>23</sup>Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.

<sup>24</sup>Estonian Biocentre, Evolutionary Biology group, Tartu 51010, Estonia. <sup>25</sup>Laboratorio de

Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE),

CCT-CONICET La Plata/CIC Buenos Aires/Universidad Nacional de La Plata, La Plata

B1906APO, Argentina. <sup>26</sup>Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.

<sup>27</sup>Department of Clinical Science, University of Bergen, Bergen 5021, Norway. <sup>28</sup>National

Laboratory of Genomics for Biodiversity (LANGEBIO), CINVESTAV, Irapuato, Guanajuato

36821, Mexico. <sup>29</sup>Institute of Cytology and Genetics, Siberian Branch of Russian Academy

of Sciences, Novosibirsk 630090, Russia. <sup>30</sup>Novosibirsk State University, Novosibirsk

630090, Russia. <sup>31</sup>Research Centre for Medical Genetics, Moscow 115478, Russia. <sup>32</sup>Vavilov

Institute for General Genetics, Moscow 119991, Russia. <sup>33</sup>Moscow Institute for Physics and

Technology, Dolgoprudniy 141700, Russia. <sup>34</sup>Department of Medical Genetics, National

Human Genome Center, Medical University Sofia, Sofia 1431, Bulgaria. <sup>35</sup>Laboratory of

Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences of Armenia,

Yerevan 0014, Armenia. <sup>36</sup>The Wellcome Trust Sanger Institute, Wellcome Genome Campus,

Hinxton, Cambridgeshire CB10 1SA, UK. <sup>37</sup>RIPAS Hospital, Bandar Seri Begawan, Brunei.

<sup>38</sup>Department of Genetics, Evolution and Environment, University College London WC1E

6BT, UK. <sup>39</sup>Department of Anthropology, Case Western Reserve University, Cleveland, Ohio

44106-7125, USA. <sup>40</sup>Laboratory of Human Molecular Genetics, Institute of Molecular and

Cellular Biology, Siberian Branch of Russian Academy of Sciences, Novosibirsk 630090,

Russia. <sup>41</sup>Department of Evolutionary Biology, University of Tartu, Tartu 51010, Estonia.

<sup>42</sup>Estonian Academy of Sciences, Tallinn 10130, Estonia. <sup>43</sup>Department of Ecology and

Evolution, Stony Brook University, Stony Brook, New York 11794, USA. <sup>44</sup>NextBio, Illumina,

Santa Clara, California 95050, USA. <sup>45</sup>Gladstone Institutes, San Francisco, California 94158,

USA. <sup>46</sup>Department of Forensic Medicine, University of Helsinki, Helsinki 00014, Finland.

<sup>47</sup>Department of Medicine, Division of Medical Genetics, University of Washington, Seattle,

Washington 98195, USA. <sup>48</sup>National Cancer Centre Singapore, 169610 Singapore. <sup>49</sup>Institute

of Biochemistry and Genetics, Ufa Research Centre, Russian Academy of Sciences, Ufa

450054, Russia. <sup>50</sup>Department of Genetics and Fundamental Medicine, Bashkir State

University, Ufa 450074, Russia. <sup>51</sup>Jaramogi Oginga Odinga University of Science and

Technology, Bondo 40601, Kenya. <sup>52</sup>Institut de Biologia Evolutiva (CSIC-UPF), Departament

de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona 08003,

Spain. <sup>53</sup>ARL Division of Biotechnology, University of Arizona, Tucson, Arizona 85721, USA.

<sup>54</sup>Division of Biological Anthropology, University of Cambridge, Fitzwilliam Street, Cambridge

CB2 1QH, UK. <sup>55</sup>Basic Research Laboratory, Center for Cancer Research, NCI, Leidos

Biomedical Research, Inc., Frederick National Laboratory, Frederick, Maryland 21702, USA.

<sup>56</sup>CHU Sainte-Justine, Pediatrics Department, Université de Montréal, Québec H3T 1C5,

Canada. <sup>57</sup>Department of Pediatrics, University of Washington, Seattle, Washington 98119,

USA. <sup>58</sup>Department of Human Genetics, University of Utah School of Medicine, Salt Lake

City, Utah 84112, USA. <sup>59</sup>Departments of Genetics and Biology, University of Pennsylvania,

Philadelphia, Pennsylvania 19104, USA. <sup>60</sup>Department of Human Genetics, Eccles Institute

of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA. <sup>61</sup>Department

of Paleolithic Archaeology, Institute of Archaeology and Ethnography, Siberian Branch

of Russian Academy of Sciences, Novosibirsk 630090, Russia. <sup>62</sup>Altai State University, Barnaul

656000, Russia. <sup>63</sup>CSIR-Centre for Cellular and Molecular Biology, Hyderabad 500 007,

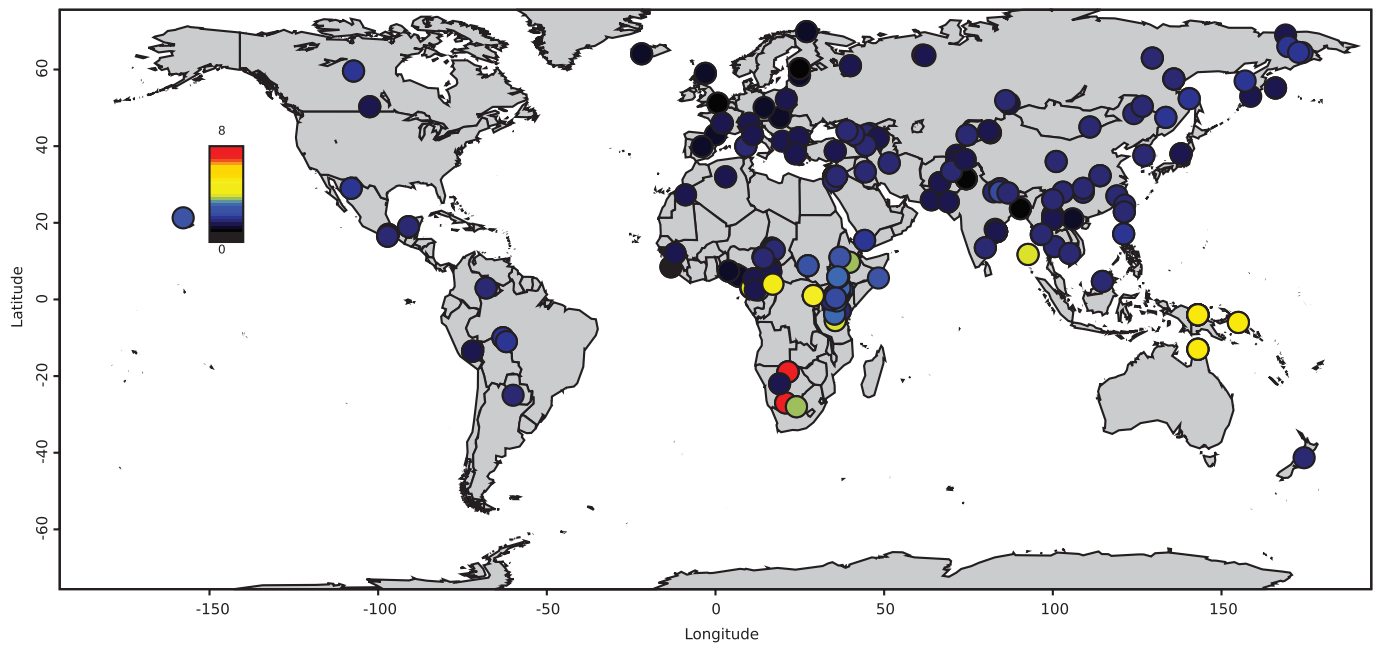
India. †Present addresses: Department of Computer Science, University of California at Los

Angeles, California 90095, USA and Department of Human Genetics Science, University of

California at Los Angeles, California 90095, USA (S.S.); Genome Foundation, Hyderabad

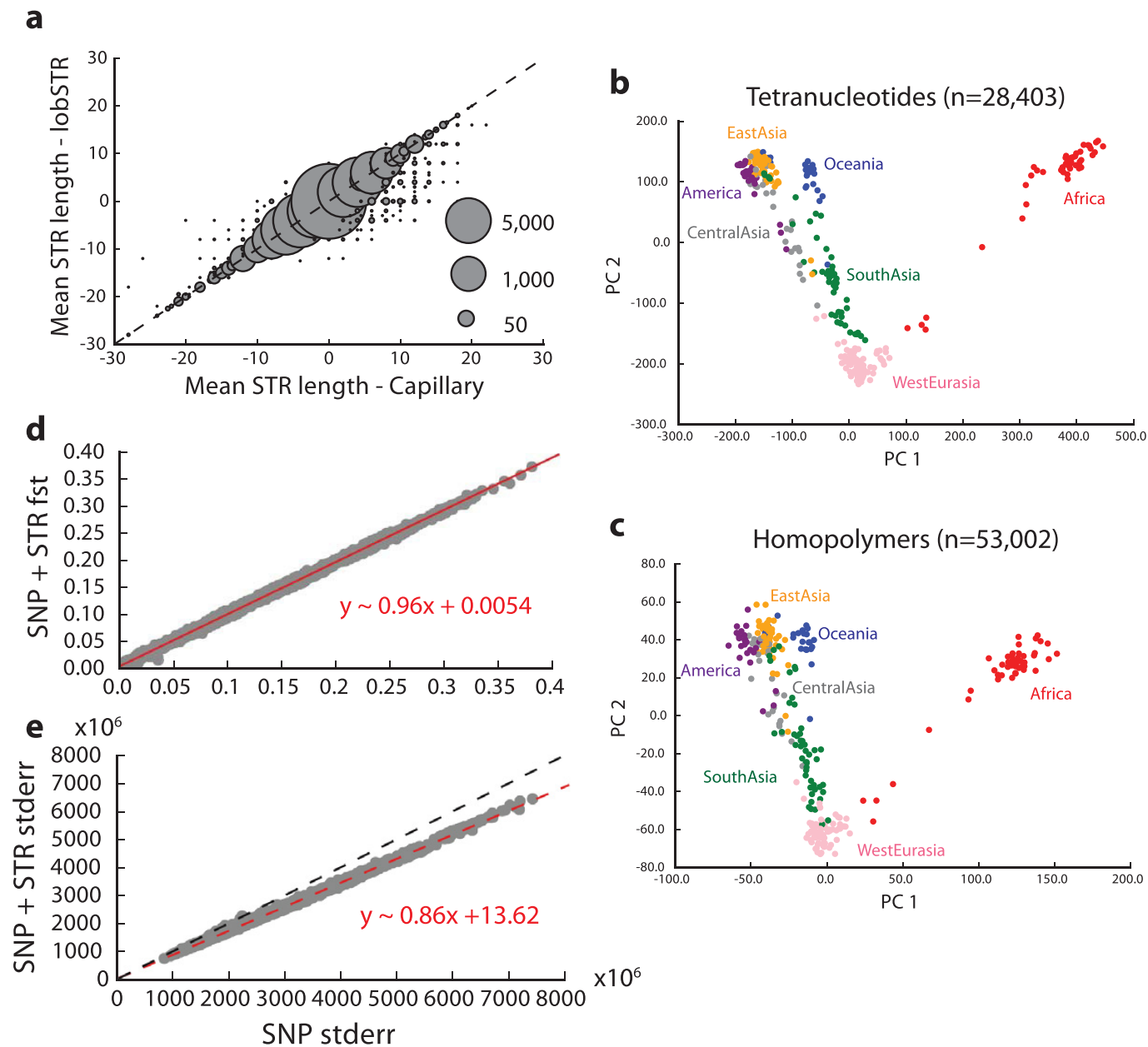
500076, India (L.S.).

\*These authors contributed equally to this work.



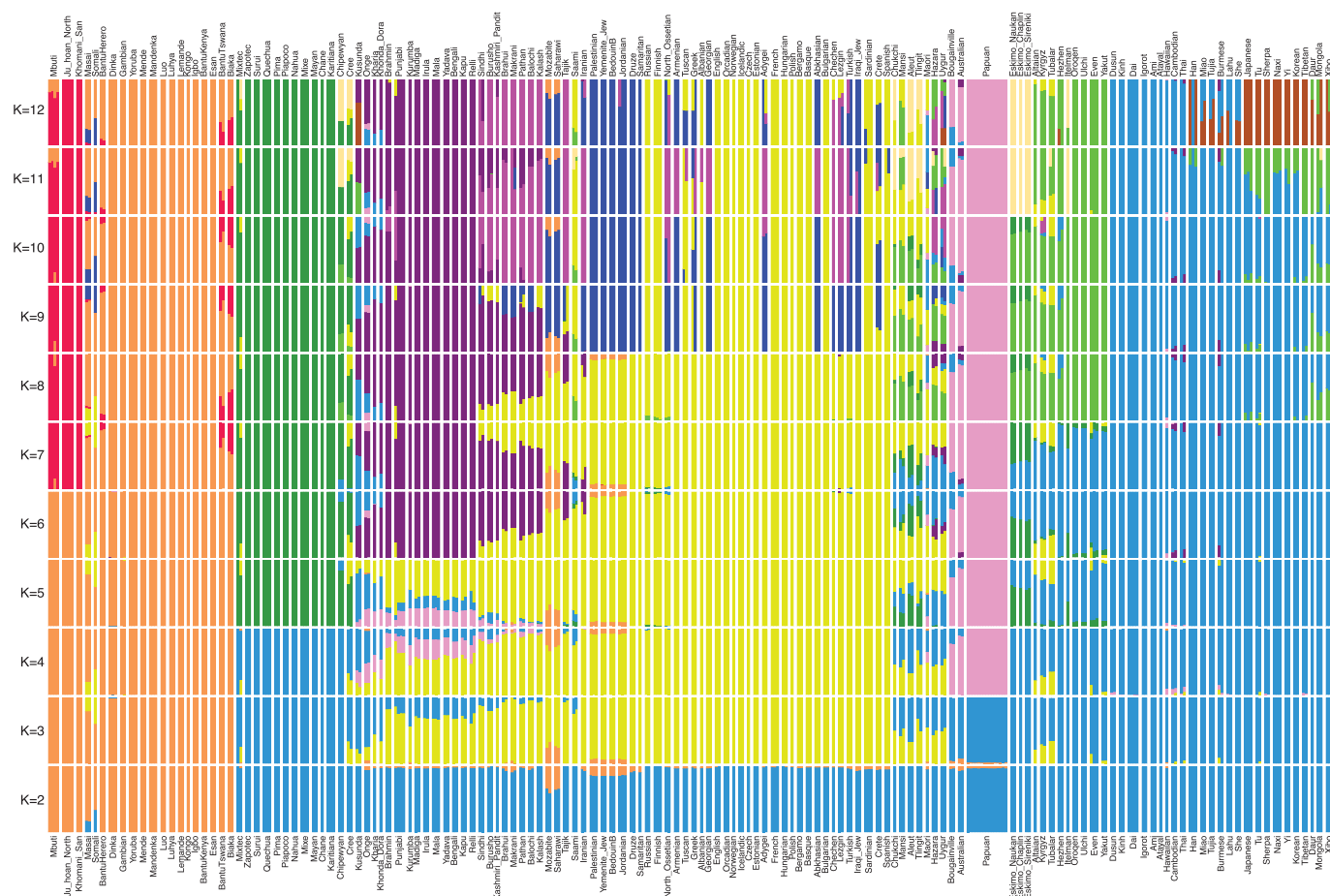
**Extended Data Figure 1 | Heat map of fraction of heterozygous sites missed in the 1000 Genomes project.** For each sample, we examine all heterozygous sites passing filter level 1, and compute the fraction included as known polymorphisms in the 1000 Genomes project.





**Extended Data Figure 2 | Worldwide variation in human short tandem repeats.** **a**, Mean STR length is reported as the average of the length difference (in base pairs) from the GRCh37 reference for each genotype. Bubble area scales with the number of calls compared at each point. **b**, **c**, The first two principal components after performing principal component analysis on tetranucleotide and homopolymer genotypes,

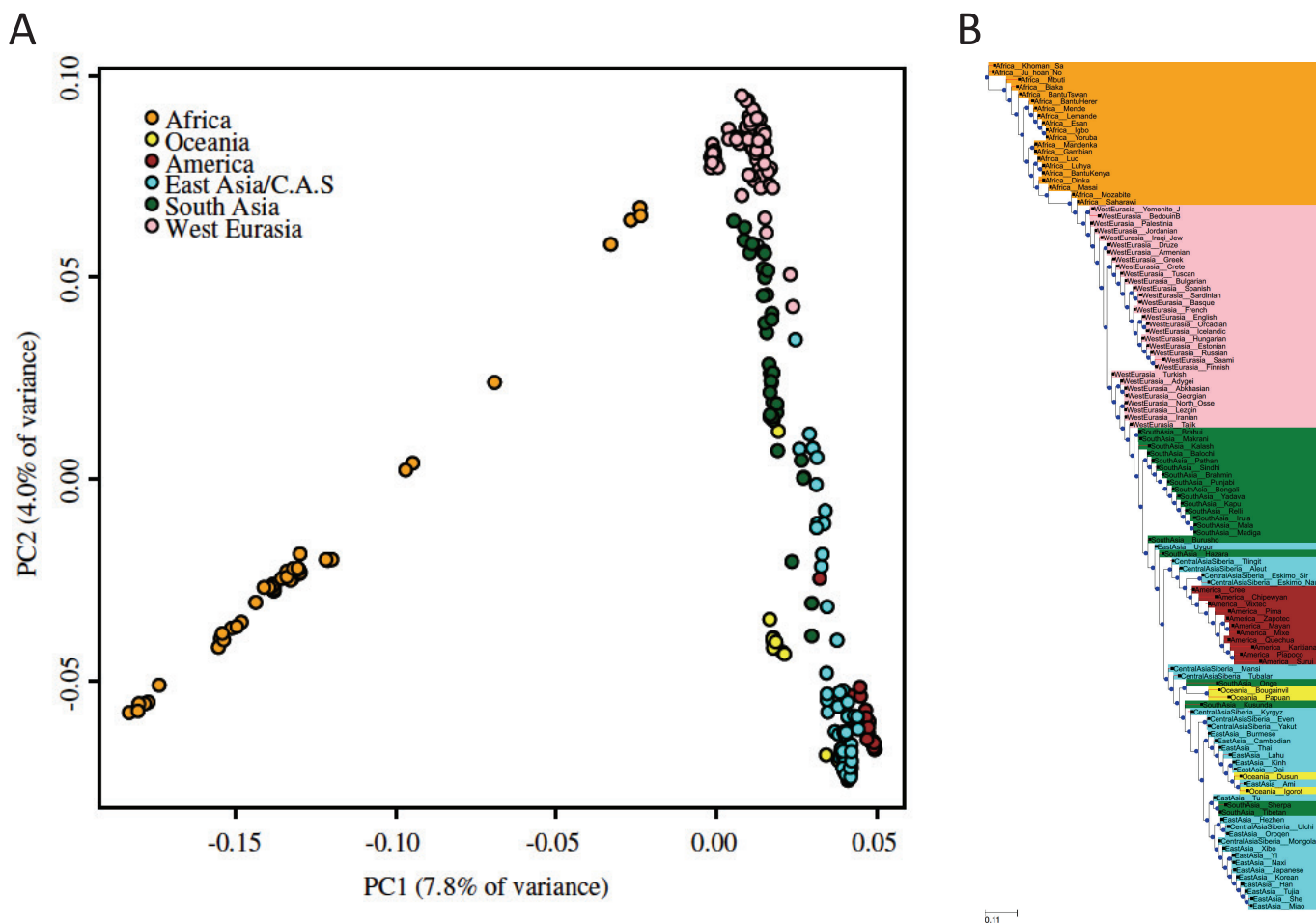
respectively. Colours represent the region of origin of each sample. **d**, Pairwise  $F_{ST}$  values between populations computed using only SNPs versus using combined SNP + STR loci. **e**, Block jackknife standard errors for the SNP versus SNP + STR  $F_{ST}$  analysis. The red dashed lines give the best-fit line, described by the formula in red. The black dashed line denotes the diagonal.



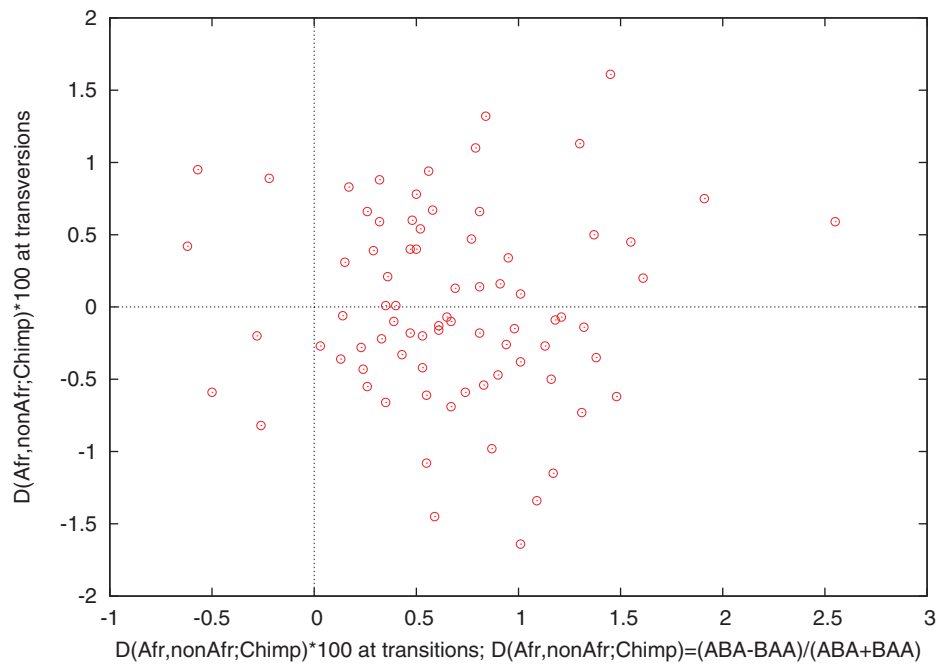
**Extended Data Figure 3 | ADMIXTURE analysis.** We carried out unsupervised ADMIXTURE 1.23<sup>8,43</sup> analysis over the 300 SGDP individuals in 20 replicates with randomly chosen initial seeds, varying the number of ancestral populations between  $K = 2$  and  $K = 12$  and using default fivefold cross-validation ( $-cv$  flag). We used genotypes of at least filter level 1, and restricted analysis to sites where at least two individuals carried the variant allele (as singleton variants are non-informative for population clustering). After further filtering of sites with at least 99% completeness and performing linkage-disequilibrium-based pruning

in PLINK 1.9<sup>44,45</sup> with parameters ( $-indep-pairwise$  1000 100 0.2), a total of 482,515 single nucleotide polymorphisms remained. This figure shows the highest likelihood replicate for each value of  $K$ . We found that log likelihood monotonically increases with  $K$ , while the value  $K = 5$  minimizes cross-validation error (not shown). The solution at  $K = 5$  corresponds to major continental groups (Sub-Saharan Africans, Oceanians, East Asians, Native Americans, and West Eurasians), but we show the full range of  $K$  here as they illustrate finer-scale population structure that may be useful to users of the data.





**Extended Data Figure 4 | Principal component analysis and neighbour joining tree. a,** Principal component analysis. **b,** Neighbour-joining tree based on  $F_{ST}$  values for all populations with at least two samples.

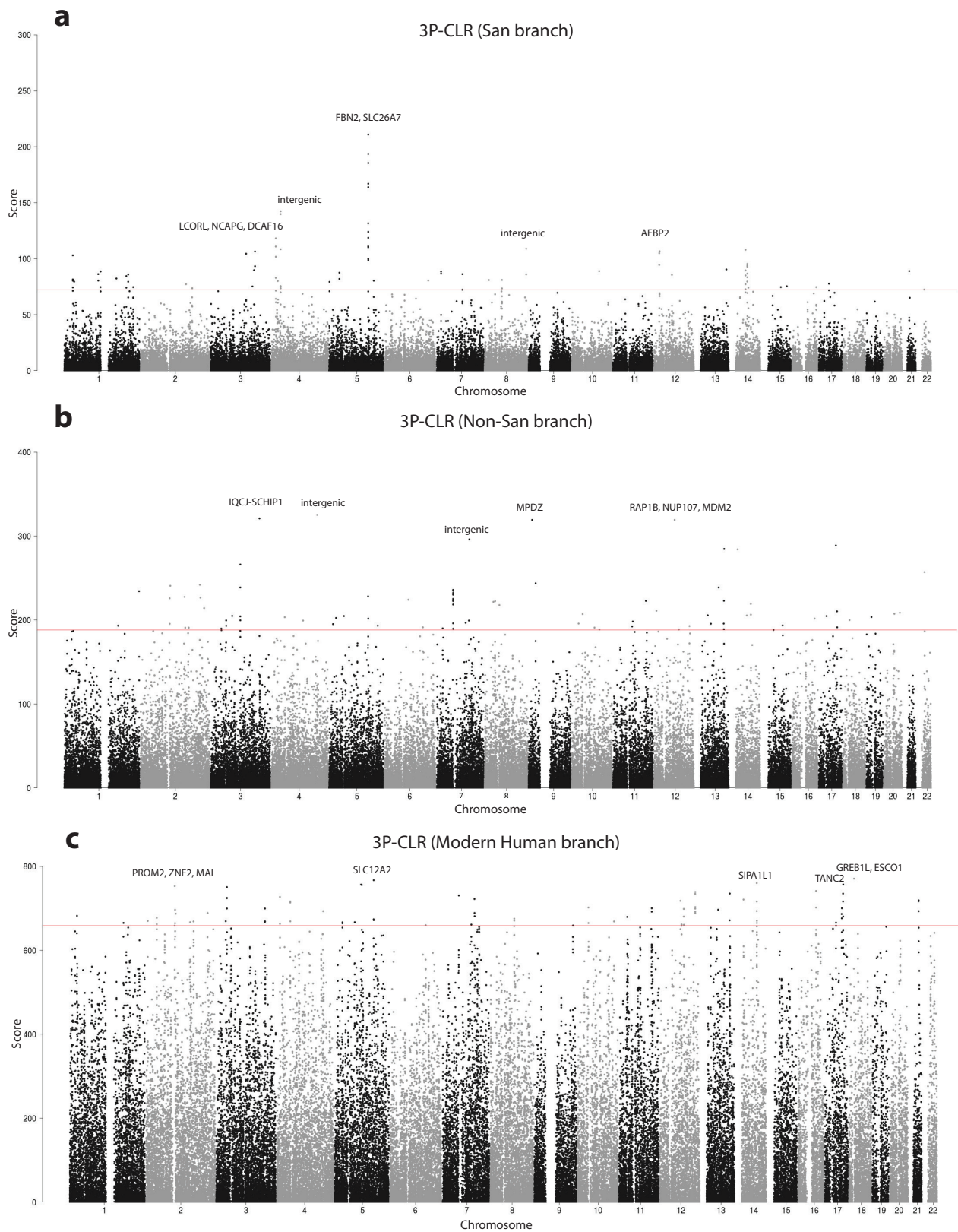


**Extended Data Figure 5 | Fewer accumulated mutations in Africans than in non-Africans confirmed by mapping to chimpanzee.**

We compute a statistic  $D$  (Population A, Population B, Chimp), measuring the difference in the rate of matching to chimpanzee in Population A compared to Population B. The evidence of mismatching to chimpanzee is seen when we restrict to the male X chromosome to eliminate possible

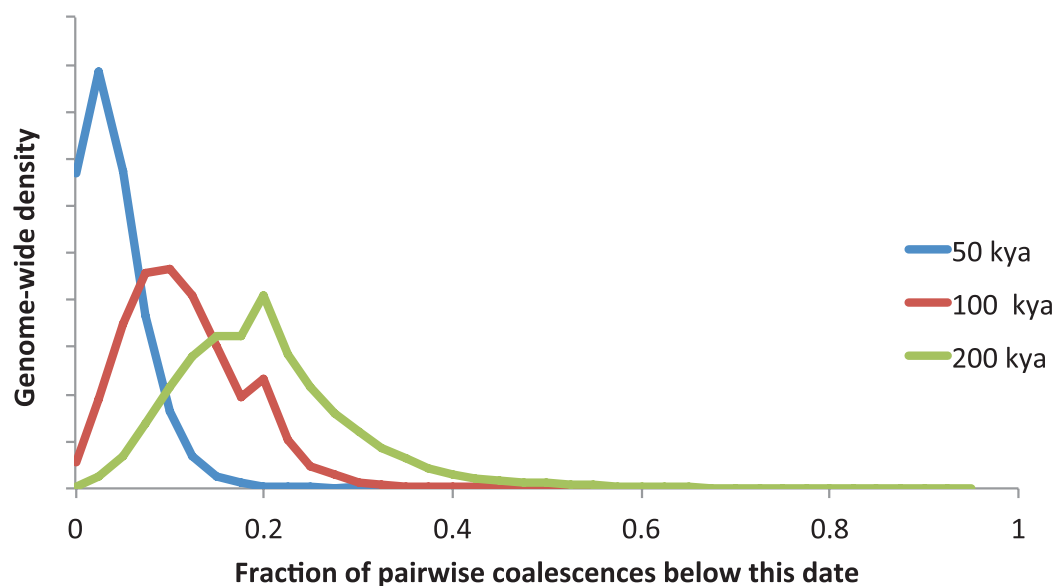
effects due to differences in heterozygosity across populations, and map to the chimpanzee genome which is phylogenetically symmetrically related to all present-day humans. We find that in 78 randomly chosen Population A = African and Population B = non-African pairs of males, transversion substitutions show no consistent skew from zero, but transition substitutions do.



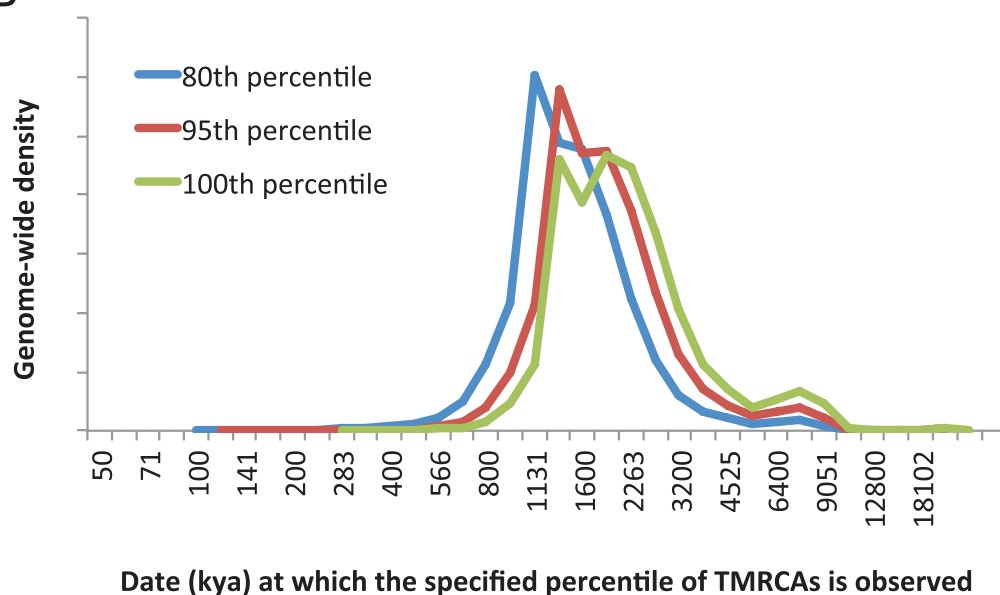


**Extended Data Figure 6 | 3P-CLR scan for positive selection.** The red line denotes the 99.9% quantile cut-off. The genes in the top five regions are labelled. **a**, Scan for selection on the San terminal branch. **b**, Scan for selection on the non-San terminal branch. **c**, Scan for selection on the ancestral modern human branch.

A



B



C

Date (in kya) at which a specified fraction X of loci have fraction Y of TMRCAs below date

X=Fraction of loci below threshold	Y=80% of TMRCAs less than this date	Y=95% of TMRCAs less than this date	Y=100% of TMRCAs less than this date
0.01%	120 kya	300 kya	430 kya
0.1%	320 kya	500 kya	620 kya
1%	580 kya	810 kya	980 kya

**Extended Data Figure 7 | Scan for genomic locations where the great majority of present-day humans share a recent common ancestor.** We carried out PSMC analysis on 40 pairs of haploid genomes chosen to sample some of the most deeply divergent present-day human lineages. We recorded the time since the most recent common ancestor (TMRCAs) at each position, and rescaled to obtain an estimate of absolute time

(Supplementary Information section 12). **a**, Distribution across the genome of the fraction of TMRCAs below specified date cut-offs. For the 100 kya cut-off, the maximum fraction observed anywhere in the genome is 68%. **b**, Distribution across the genome of the date  $T$  at which specified fractions of sample pairs are inferred to have a TMRCAs less than  $T$ . **c**, Percentile points of the cumulative distribution function of **B**.



Extended Data Table 1 | Fewer accumulated mutations in Africans than in non-Africans

Population A	Population B	All autosomes		All X chromosome		Lowest B quintile		Highest B quintile	
		D×100	Z	D×100	Z	D×100	Z	D×100	Z
Khoesan	Oceania	-0.35	-8.2	-0.70	-2.7	-0.68	-6.4	-0.14	-1.7
Africa	America	-0.33	-9.4	-0.73	-2.8	-0.65	-7.3	-0.18	-2.6
Khoesan	WestEurasia	-0.30	-7.5	-0.68	-3.1	-0.63	-6.3	-0.17	-2.1
Africa	Oceania	-0.29	-8.5	-0.66	-3.2	-0.55	-6.6	-0.07	-1.0
Africa	WestEurasia	-0.25	-8.5	-0.66	-3.1	-0.49	-6.4	-0.11	-1.8
Khoesan	SouthAsia	-0.24	-6.0	-0.56	-2.7	-0.61	-6.3	-0.11	-1.4
Africa	EastAsia	-0.20	-6.6	-0.65	-2.5	-0.42	-5.2	-0.10	-1.5
Africa	CentralAsiaSiberia	-0.20	-6.2	-0.55	-2.2	-0.48	-6.3	-0.05	-0.7
Pygmy	WestEurasia	-0.19	-4.8	-0.46	-1.4	-0.43	-4.6	-0.04	-0.5
Africa	SouthAsia	-0.18	-6.4	-0.50	-2.0	-0.46	-6.3	-0.03	-0.5
CentralAsiaSiberia	Oceania	-0.13	-3.9	-0.15	-0.6	-0.09	-1.1	-0.03	-0.4
Pygmy	SouthAsia	-0.13	-3.3	-0.38	-1.1	-0.38	-4.2	0.02	0.2
EastAsia	Oceania	-0.13	-4.1	0.00	0.0	-0.17	-2.1	0.04	0.6
Khoesan	Pygmy	-0.10	-2.6	-0.14	-0.4	-0.16	-1.6	-0.12	-1.5
SouthAsia	WestEurasia	-0.08	-4.3	-0.20	-1.2	-0.05	-1.0	-0.10	-2.7
CentralAsiaSiberia	WestEurasia	-0.06	-2.2	-0.16	-0.8	-0.01	-0.2	-0.09	-1.6
EastAsia	WestEurasia	-0.06	-2.1	-0.00	-0.0	-0.08	-1.0	-0.02	-0.3
CentralAsiaSiberia	EastAsia	-0.00	-0.2	-0.18	-1.1	0.07	1.2	-0.08	-1.8
Africa	Pygmy	-0.00	-0.1	-0.06	-0.2	0.03	0.4	-0.06	-0.8
EastAsia	SouthAsia	0.02	0.7	0.22	1.7	-0.04	-0.7	0.08	1.7
CentralAsiaSiberia	SouthAsia	0.02	0.7	0.05	0.3	0.02	0.4	-0.00	-0.0
America	Oceania	0.03	0.9	0.11	0.4	0.10	1.1	0.13	1.7
Oceania	WestEurasia	0.08	2.3	-0.03	-0.1	0.10	1.1	-0.04	-0.6
Africa	Khoesan	0.10	2.9	0.17	0.7	0.23	2.6	0.07	1.0
America	WestEurasia	0.11	3.6	0.11	0.4	0.19	2.2	0.08	1.3
CentralAsiaSiberia	Pygmy	0.14	3.4	0.32	0.9	0.43	4.5	-0.04	-0.4
Oceania	SouthAsia	0.14	4.8	0.22	0.9	0.13	1.7	0.04	0.7
EastAsia	Pygmy	0.15	3.6	0.49	1.4	0.37	3.9	0.04	0.5
America	EastAsia	0.18	5.9	0.09	0.3	0.28	3.6	0.11	1.8
America	CentralAsiaSiberia	0.18	6.2	0.34	1.7	0.23	2.9	0.18	3.1
America	SouthAsia	0.18	6.4	0.34	1.5	0.22	3.0	0.18	3.1
Oceania	Pygmy	0.24	5.4	0.46	1.3	0.45	4.6	0.02	0.2
CentralAsiaSiberia	Khoesan	0.25	6.0	0.57	2.9	0.64	6.3	0.09	1.1
EastAsia	Khoesan	0.25	6.2	0.68	3.2	0.59	5.9	0.14	1.7
America	Pygmy	0.26	5.9	0.58	1.6	0.58	5.7	0.09	1.0
America	Khoesan	0.37	8.7	0.76	3.3	0.77	7.3	0.22	2.5

We compute a statistic  $D$  (Population A, Population B, Chimp), measuring the difference in the rate of matching to chimpanzee in Population A compared to Population B. For all the autosomes, we observe highly significant signals ( $3.3 < |Z| < 9.4$ ) of excess mismatching to chimpanzee in non-Africans compared to Africans, using a standard error from a block jack-knife. We highlight  $|D| > 0.002$  in blue, and  $|Z| > 3$  in yellow. The deviations from zero are greatest in subsets of the genome where the time since two populations split comprises a relatively larger fraction of the total genetic divergence time between the populations; this is the direction expected from a mutation accumulation change since divergence. Compared to all the autosomes as a baseline, a least squares fit indicate that the deviations are 2.2 times higher on chromosome X, 2.0 times higher in the quintile of lowest B-statistic (closest to functionally important regions), and 0.43 times as high in the quintile of lowest B-statistic (furthest from functional regions).

# A genomic history of Aboriginal Australia

Anna-Sapfo Malaspinas<sup>1,2,3\*</sup>, Michael C. Westaway<sup>4\*</sup>, Craig Muller<sup>1\*</sup>, Vitor C. Sousa<sup>2,3\*</sup>, Oscar Lao<sup>5,6\*</sup>, Isabel Alves<sup>2,3,7\*</sup>, Anders Bergström<sup>8\*</sup>, Georgios Athanasiadis<sup>9</sup>, Jade Y. Cheng<sup>9,10</sup>, Jacob E. Crawford<sup>10,11</sup>, Tim H. Heupink<sup>4</sup>, Enrico Macholdt<sup>12</sup>, Stephan Peischl<sup>3,13</sup>, Simon Rasmussen<sup>14</sup>, Stephan Schiffels<sup>15</sup>, Sankar Subramanian<sup>4</sup>, Joanne L. Wright<sup>4</sup>, Anders Albrechtsen<sup>16</sup>, Chiara Barbieri<sup>12,17</sup>, Isabelle Dupanloup<sup>2,3</sup>, Anders Eriksson<sup>18,19</sup>, Ashot Margaryan<sup>1</sup>, Ida Moltke<sup>16</sup>, Irina Pugach<sup>12</sup>, Thorfinn S. Korneliussen<sup>1</sup>, Ivan P. Levkivskyi<sup>20</sup>, J. Víctor Moreno-Mayar<sup>1</sup>, Shengyu Ni<sup>12</sup>, Fernando Racimo<sup>10</sup>, Martin Sikora<sup>1</sup>, Yali Xue<sup>8</sup>, Farhang A. Aghakhanian<sup>21</sup>, Nicolas Brucato<sup>22</sup>, Søren Brunak<sup>23</sup>, Paula F. Campos<sup>1,24</sup>, Warren Clark<sup>25</sup>, Sturla Ellingvåg<sup>26</sup>, Gudjugudju Fourmile<sup>27</sup>, Pascale Gerbault<sup>28,29</sup>, Darren Injie<sup>30</sup>, George Koki<sup>31</sup>, Matthew Leavesley<sup>32</sup>, Betty Logan<sup>33</sup>, Aubrey Lynch<sup>34</sup>, Elizabeth A. Matisoo-Smith<sup>35</sup>, Peter J. McAllister<sup>36</sup>, Alexander J. Mentzer<sup>37</sup>, Mait Metspalu<sup>38</sup>, Andrea B. Migliano<sup>29</sup>, Les Murgu<sup>39</sup>, Maude E. Phipps<sup>21</sup>, William Pomat<sup>31</sup>, Doc Reynolds<sup>40</sup>, Francois-Xavier Ricaut<sup>22</sup>, Peter Siba<sup>31</sup>, Mark G. Thomas<sup>28</sup>, Thomas Wales<sup>41</sup>, Colleen Ma'run Wall<sup>42</sup>, Stephen J. Oppenheimer<sup>43</sup>, Chris Tyler-Smith<sup>8</sup>, Richard Durbin<sup>8</sup>, Joe Dortch<sup>44</sup>, Andrea Manica<sup>18</sup>, Mikkel H. Schierup<sup>9</sup>, Robert A. Foley<sup>1,45</sup>, Marta Mirazón Lahr<sup>1,45</sup>, Claire Bowern<sup>46</sup>, Jeffrey D. Wall<sup>47</sup>, Thomas Mailund<sup>9</sup>, Mark Stoneking<sup>12</sup>, Rasmus Nielsen<sup>1,48</sup>, Manjinder S. Sandhu<sup>8</sup>, Laurent Excoffier<sup>2,3</sup>, David M. Lambert<sup>4</sup> & Eske Willerslev<sup>1,8,18</sup>

**The population history of Aboriginal Australians remains largely uncharacterized. Here we generate high-coverage genomes for 83 Aboriginal Australians (speakers of Pama-Nyungan languages) and 25 Papuans from the New Guinea Highlands. We find that Papuan and Aboriginal Australian ancestors diversified 25–40 thousand years ago (kya), suggesting pre-Holocene population structure in the ancient continent of Sahul (Australia, New Guinea and Tasmania). However, all of the studied Aboriginal Australians descend from a single founding population that differentiated ~10–32 kya. We infer a population expansion in northeast Australia during the Holocene epoch (past 10,000 years) associated with limited gene flow from this region to the rest of Australia, consistent with the spread of the Pama-Nyungan languages. We estimate that Aboriginal Australians and Papuans diverged from Eurasians 51–72 kya, following a single out-of-Africa dispersal, and subsequently admixed with archaic populations. Finally, we report evidence of selection in Aboriginal Australians potentially associated with living in the desert.**

During most of the last 100,000 years, Australia, Tasmania and New Guinea formed a single continent, Sahul, which was separated from Sunda (the continental landmass including mainland and western island Southeast Asia) by a series of deep oceanic troughs never exposed by changes in sea level. Colonization of Sahul is thought to have required at least 8–10 sea crossings between islands, potentially constraining the occupation of Australia and New Guinea by

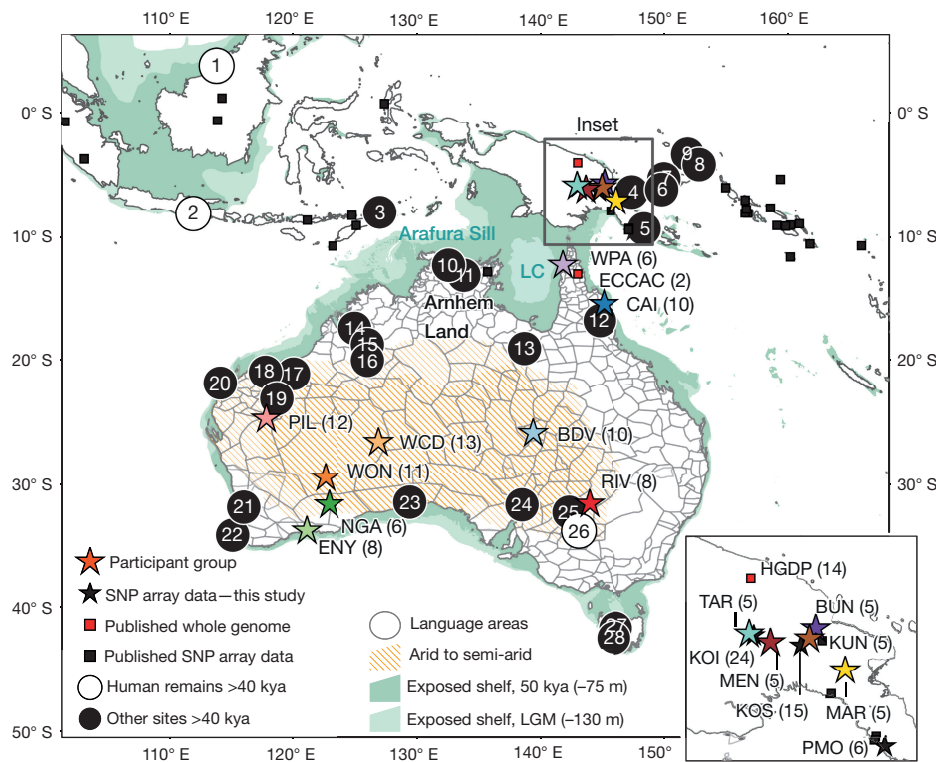
earlier hominins<sup>1</sup>. Recent assessments suggest that Sahul was settled by 47–55 kya<sup>2,3</sup> (Fig. 1). These dates align with those for the earliest evidence for modern humans in Sunda<sup>4</sup>.

The distinctiveness of the Australian archaeological and fossil record has led to the suggestion that the ancestors of Aboriginal Australians and Papuans ('Australo-Papuans' hereafter) left the African continent earlier than the ancestors of present-day Eurasians<sup>5</sup>. Although some

<sup>1</sup>Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen, Denmark. <sup>2</sup>Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland. <sup>3</sup>Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. <sup>4</sup>Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Nathan, Queensland 4111, Australia. <sup>5</sup>CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain. <sup>6</sup>Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. <sup>7</sup>Population and Conservation Genetics Group, Instituto Gulbenkian de Ciência, 2780-156 Oeiras, Portugal. <sup>8</sup>Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. <sup>9</sup>Bioinformatics Research Centre, Aarhus University, 8000 Aarhus, Denmark. <sup>10</sup>Department of Integrative Biology, University of California, Berkeley, California 94720, USA. <sup>11</sup>Verily Life Sciences, 2425 Garcia Ave, Mountain View, California 94043, USA. <sup>12</sup>Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. <sup>13</sup>Interfaculty Bioinformatics Unit University of Bern, Baltzerstrasse 6, CH-3012 Bern, Switzerland. <sup>14</sup>Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, 2800 Kongens Lyngby, Denmark. <sup>15</sup>Department for Archaeogenetics, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, D-07745 Jena, Germany. <sup>16</sup>The Bioinformatics Centre, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark. <sup>17</sup>Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Kahlaische Straße 10, D-07745 Jena, Germany. <sup>18</sup>Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK. <sup>19</sup>Integrative Systems Biology Laboratory, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, 23955-6900 Thuwal, Saudi Arabia. <sup>20</sup>Institute for Theoretical Physics, ETH Zürich, Wolfgang-Pauli-Str. 27, 8093 Zürich, Switzerland. <sup>21</sup>Jeffrey Cheah School of Medicine & Health Sciences, Monash University Malaysia, Jalan Lagoon Selatan, Sunway City, 46150 Selangor, Malaysia. <sup>22</sup>Evolutionary Medicine Group, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse, UMR 5288, Centre National de la Recherche Scientifique, Université de Toulouse 3, 31073 Toulouse, France. <sup>23</sup>Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen N, Denmark. <sup>24</sup>CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua das Bragas 289, 4050-123 Porto, Portugal. <sup>25</sup>National Parks and Wildlife, Sturt Highway, Buronga, New South Wales 2739, Australia. <sup>26</sup>Explico Foundation, Vågavegen 16, 6900 Florø, Norway. <sup>27</sup>Giruwandi, Gimuy Yidinji Country, Queensland 4868, Australia. <sup>28</sup>Research Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK. <sup>29</sup>UCL Department of Anthropology, 14 Taviston Street, London WC1H 0BW, UK. <sup>30</sup>Yinhawangka elder, Perth, Western Australia 6062, Australia. <sup>31</sup>Papua New Guinea Institute of Medical Research, PO Box 60, Goroka, Papua New Guinea. <sup>32</sup>Archaeology, School of Humanities & Social Sciences, University PO Box 320, University of Papua New Guinea & College of Arts, Society & Education, James Cook University, Cairns, Queensland 4811, Australia. <sup>33</sup>Ngadjju elder, Coolgardie, Western Australia 6429, Australia. <sup>34</sup>Wongatha elder, Kurrurang, Western Australia 6430, Australia. <sup>35</sup>Department of Anatomy, University of Otago, Dunedin 9054, New Zealand. <sup>36</sup>2209 Springbrook Road, Springbrook, Queensland 4213, Australia. <sup>37</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. <sup>38</sup>Estonian Biocentre, Riia 23b, Tartu 51010, Estonia. <sup>39</sup>86 Workshop Road, Yarrabah, Queensland 4871, Australia. <sup>40</sup>Esperance Nyungar elder, Esperance, Western Australia 6450, Australia. <sup>41</sup>Atakani Street, Napranum, Queensland 4874, Australia. <sup>42</sup>Wynnum North Road, Wynnum, Queensland 4178, Australia. <sup>43</sup>School of Anthropology and Museum Ethnography, Oxford University, Oxford OX2 6PE, UK. <sup>44</sup>Centre for Rock Art Research and Management, M257, University of Western Australia, Perth, Western Australia 6009, Australia. <sup>45</sup>Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Fitzwilliam Street, Cambridge CB2 1QH, UK. <sup>46</sup>Department of Linguistics, Yale University, 370 Temple Street, New Haven, Connecticut 06520, USA. <sup>47</sup>Institute for Human Genetics, University of California, San Francisco, California 94143, USA. <sup>48</sup>Departments of Integrative Biology and Statistics, University of California, Berkeley, California 94720, USA.

\*These authors contributed equally to this work.





**Figure 1 | Aboriginal Australian and Papuan samples used in this study, as well as archaeological sites and human remains dated to ~40 kya or older in southern Sunda and Sahul.** The stars indicate the centroid location for each sampling group (sample size in parentheses). Publicly available genetic data (see Supplementary Information section S04) used as a reference panel in this study are shown as squares. Sites with dated human remains are shown as white circles and the archaeological sites as black circles. The associated dates can be found in Supplementary Information section S03. Grey boundaries correspond to territories defined by the language groups provided by the Australian Institute of Aboriginal and Torres Strait Islander Studies<sup>45</sup>. Sampled Aboriginal Australians self-identify primarily as: Yidindji and Gungandji from the Cairns region (CAI,  $n = 10$ , see also Supplementary Information section S02); Yupangati and Thanakwithi from northwest Cape York (WPA,  $n = 6$ ), Wangkangurru and Yarluyandi from the Birdsville region (BDV,

$n = 10$ , 9 sequenced at high depth), Barkindji from southeast (RIV,  $n = 8$ ); Pilbara area Yinhawangka and Banjima (PIL,  $n = 12$ ), Ngaanyatjarra from western central desert (WCD,  $n = 13$ ), Wongatha from Western Australia's northern Goldfields (WON,  $n = 11$ ), Ngadjju from Western Australia's southern Goldfields (NGA,  $n = 6$ ); and Nyungar from southwest Australia (ENY,  $n = 8$ ). Papuans include samples from the locations Bundi (BUN,  $n = 5$ ), Kundiawa (KUN,  $n = 5$ ), Mendi (MEN,  $n = 5$ ), Marawaka (MAR,  $n = 5$ ) and Tari (TAR,  $n = 5$ ). We generated SNP array data (black stars) for 45 Papuan samples including 24 Koinambe (KOI) and 15 Kosipe (KOS)—described previously<sup>46</sup>—and 6 individuals with Highland ancestry sampled in Port Moresby (PMO). Lake Carpentaria (LC), which covered a significant portion of the land bridge between Australia and New Guinea 11.5–40 kya and thus potentially acted as a barrier to gene flow, is also indicated. Map data were sourced from the Australian Government, <http://www.natureearthdata.com/> and our research.

genetic studies support such multiple dispersals from Africa<sup>6</sup>, others favour only one out-of-Africa (OoA) event, with one or two independent founding waves into Asia, of which the earlier contributed to Australo-Papuan ancestry<sup>7,8</sup>. In addition, recent genomic studies have shown that both Aboriginal Australian<sup>8</sup> and Papuan<sup>9</sup> ancestors admixed with Neanderthal and Denisovan archaic hominins after leaving Africa.

Increased desertification of Australia<sup>10</sup> during the last glacial maximum (LGM) 19–26.5 kya affected the number and density of human populations<sup>11</sup>. In this context, unique morphological and physiological adaptations have been identified in Aboriginal Australians living in desert areas today<sup>12</sup>. In particular, desert groups were hypothesized to withstand sub-zero night temperatures without showing the increase in metabolic rates observed in Europeans under the same conditions.

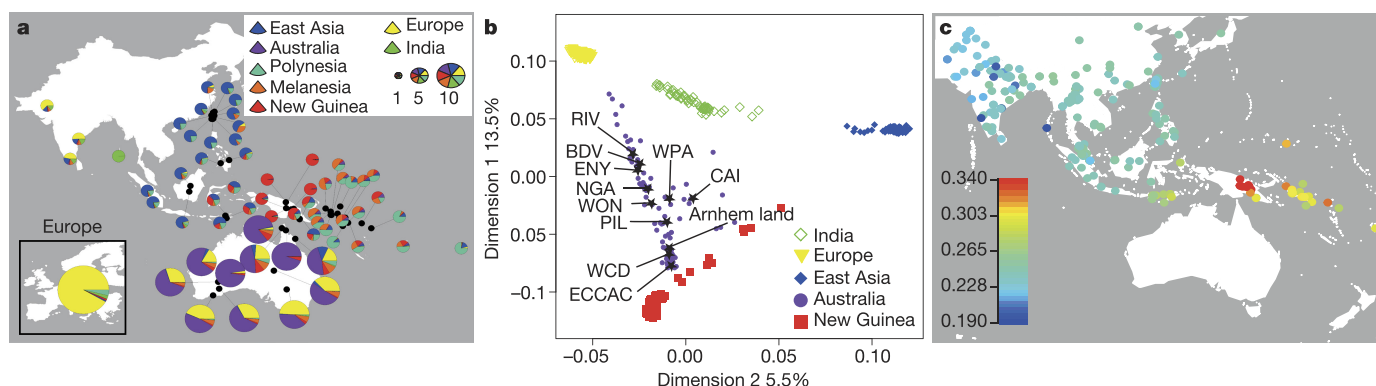
At the time of European contact, Aboriginal Australians spoke over 250 distinct languages, two-thirds of which belong to the Pama–Nyungan family and cover 90% of the Australian mainland<sup>13</sup>. The place of origin of this language family and the effect of its extensive diffusion on its internal phylogenetic structure have been debated<sup>14</sup>, but the pronounced similarity among Pama–Nyungan languages, together with shared socio-cultural patterns, have been interpreted as resulting from a mid-Holocene expansion<sup>15</sup>. Other changes in the mid-late Holocene (~4 kya) include the proliferation of backed blades and the introduction of the dingo<sup>16</sup>. It has been suggested that

Pama–Nyungan languages, dingoes and backed blades all reflect the same recent migration into Australia<sup>17</sup>. Although an external origin for backed blades has been rejected, dingoes were certainly introduced, most likely via island Southeast Asia<sup>16</sup>. A recent genetic study found evidence of Indian gene flow into Australia at the approximate time of these Holocene changes<sup>18</sup>, suggesting a possible association, while substantial admixture with Asians and Europeans is well documented in historical times<sup>19</sup>.

To date, only three Aboriginal Australian whole genome sequences have been described—one deriving from a historical tuft of hair from Australia's Western Desert<sup>8</sup> and two others from cell lines with limited provenance information<sup>20</sup>. In this study, we report the first extensive investigation of Aboriginal Australian genomic diversity by analysing the high-coverage genomes of 83 Pama–Nyungan-speaking Aboriginal Australians and 25 Highland Papuans.

## Dataset

We collected saliva samples for DNA sequencing in collaboration with Aboriginal Australian communities and individuals in Australia (Supplementary Information section S01). We sequenced genomes at high-depth (average of 60×, range 20–100×) from 83 Aboriginal Australian individuals widely distributed geographically and linguistically (see Fig. 1 and associated legend for the location and label for each group as well as Extended Data Table 1, Supplementary Information



**Figure 2 | Genetic ancestry of Aboriginal Australians in a worldwide context.** **a**, Estimation of genomic ancestry proportions for the best number of ancestral components ( $K=7$ ) based on Aboriginal Australian and Papuan whole-genome sequence and SNP array data from this study (see Fig. 1), and publicly available SNP array data<sup>18,26,47,57</sup> (Supplementary Information section S05). Each ancestry component has been labelled according to the geographic region showing the corresponding highest frequency. The area of each pie chart is proportional to the sample size (as depicted in the legend). The genomes of Aboriginal Australian populations are mostly a mixture of European and Aboriginal Australian ancestry components. Northern Aboriginal Australian groups (Arnhem Land, CAI, ECCAC, PIL and WPA) are also assigned to components mainly present in East Asian populations, while northeastern Aboriginal Australian groups (CAI and WPA) also show components mainly present in New Guinean populations. A background of 5% ‘Melanesian’ component is observed in all the Aboriginal Australian populations; however, this component is widely spread over the geographic area shown in this figure, being present from Taiwan to India. We detected on average 1.5% ‘Indian’ component and 1.4% ‘Polynesian’ component across the Aboriginal Australian samples, but we attribute these residual ancestry components to statistical noise as they are present in other Southeast Asian populations and are not supported by other analyses (Supplementary Information section S05). **b**, Classical Multidimensional scaling (MDS) plot of first two dimensions

based on an identity-by-state (IBS) distance matrix (based on 54,971 SNPs) between individuals from this study and worldwide populations, including publicly available data<sup>18,26,47,57</sup>. The first two dimensions explain 19% of the variance in the IBS distance matrix. Individuals are colour-coded according to sampling location, grouped into Australia (Arnhem Land, ECCAC, BDV, CAI, ENY, NGA, PIL, RIV, WCD, WON, WPA); East Asia (Cambodian, Dai, Han, Japanese, Naxi); Europe (English, French, Sardinian, Scottish, Spanish); India (Vishwabrahmin, Dravidian, Punjabi, Guaharati); and New Guinea (HGDP-Papuan, Central Province, Eastern Highlands, Gulf Province, Highlands, PMO, KOI, KOS, BUN, KUN, MEN, TAR, MAR). Stars indicate the centroid for each Aboriginal Australian group. Aboriginal Australians from this study as well as from previous studies are closest to Papuans and also show signals of admixture with Eurasians (see Supplementary Information section S05 for details). **c**, A heat map displaying outgroup  $f_3$  statistics of the form  $f_3(\text{Mbuti}; \text{WCD02}, X)$ , quantifying genetic drift shared between the putatively unadmixed individual WCD02 chosen to represent the Aboriginal Australian population, and various populations throughout the broader region for which either array genotypes or whole-genome sequencing data were publicly available or generated in this study. We used 760,116 SNPs for which WCD02 had non-missing array genotypes that overlapped with any other datasets. Standard errors, as estimated from block jack-knife resampling across the genome, were in the range 0.002–0.007.

sections S02–S04 for more information). Additionally, we sequenced 25 Highland Papuan genomes (38–53×; Supplementary Information sections S03, S04) from individuals representative of five linguistic groups, and generated genotype data for 45 additional Papuans living or originating in the Highlands (Fig. 1). These datasets were combined with previously published genomes and SNP array genotype data, including Aboriginal Australian data from Arnhem Land, and from a human diversity cell line panel from the European Collection of Cell Cultures<sup>20</sup> (ECCAC, Fig. 1, Supplementary Information section S04).

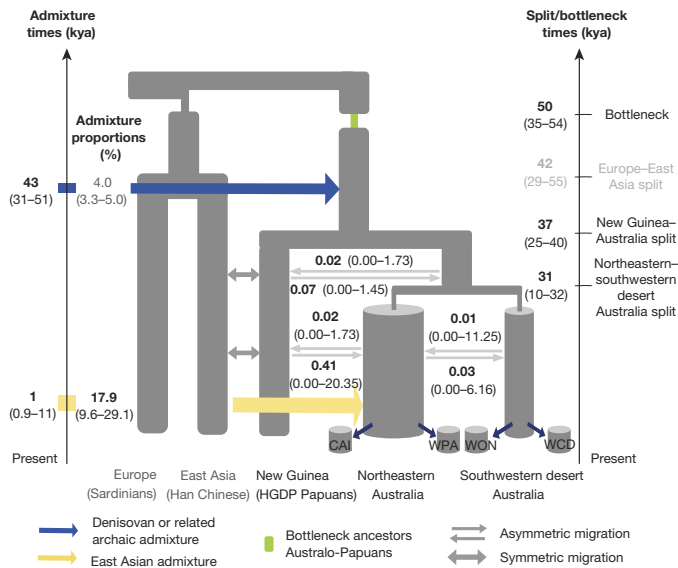
We explored the extent of admixture in the Aboriginal Australian autosomal gene pool by estimating ancestry proportions with an approach based on sparse non-negative matrix factorization (sNMF)<sup>21</sup>. We found that the genomic diversity of Aboriginal Australian populations is best modelled as a mixture of four main genetic ancestries that can be assigned to four geographic regions based on their relative frequencies: Europe, East Asia, New Guinea and Australia (Fig. 2a, Extended Data Fig. 1, Supplementary Information section S05). The degree of admixture varies among groups (Supplementary Information section S05), with the Ngaanyatjarra speakers from central Australia (WCD) having a significantly higher ‘Aboriginal Australian component’ (median value = 0.95) in their genomes than the other groups sampled (median value = 0.64; Mann–Whitney rank sum test, one-tail  $P=3.55 \times 10^{-7}$ ). The East Asian and New Guinean components are mostly present in northeastern Aboriginal Australian populations, while the European component is widely distributed across groups (Fig. 2a, Extended Data Fig. 1, Supplementary Information section S05). In most of the subsequent analyses, we either selected specific samples or groups according to their level of Aboriginal Australian ancestry, or masked the data for the non-Aboriginal Australian ancestry genomic components (Supplementary Information section S06).

## Colonization of Sahul

The origin of Aboriginal Australians is a source of much debate, as is the nature of the relationships among Aboriginal Australians, and between Aboriginal Australians and Papuans. Using  $f_3$  statistics<sup>22</sup>, estimates of genomic ancestry proportions and classical multidimensional scaling (MDS) analyses, we find that Aboriginal Australians and Papuans are closer to each other than to any other present-day worldwide population considered in our study (Fig. 2b, c, Supplementary Information section S05). This is consistent with Aboriginal Australians and Papuans originating from a common ancestral population which initially colonized Sahul. Moreover, out-group  $f_3$  statistics do not reveal any significant differences between Papuan populations (Highland Papuan groups sampled in this study and the Human Genome Diversity Project (HGDP-Papuans)) in their genetic affinities to Aboriginal Australians (Extended Data Fig. 2a), suggesting that Papuan populations diverged from one another after or at the same time as they diverged from Aboriginal Australians.

To investigate the number of founding waves into Australia, we contrasted alternative models of settlement history through a composite likelihood method that compares the observed joint site frequency spectrum (SFS) to that predicted under specific demographic models<sup>23</sup> (Fig. 3, Supplementary Information section S07). We compared HGDP-Papuans to four Aboriginal Australian population samples with low levels of European admixture (Extended Data Fig. 1) from both northeastern (CAI and WPA) and southwestern desert (WON and WCD) Australia. We compared one- and two-wave models, where each Australian region was either colonized independently, or by descendants of a single Australian founding population after its divergence from Papuans. The one-wave model provides a better fit to the





**Figure 3 | Settlement of Australia.** Best supported demographic model of the colonization of Australia and New Guinea. The demographic history of Aboriginal Australian populations was modelled by considering that sampled individuals are from sub-populations ('islands') that are part of two larger regions ('continents'), which geographically match the northeast and the southwestern desert regions of Australia. Maximum likelihood parameter estimates were obtained from the joint SFS of Han Chinese, HGDP-Papuans, CAI, WPA, WON and WCD. The 95% CI, obtained by non-parametric block bootstrap, are shown within parentheses. Estimated migration rates scaled by the effective population size ( $2Nm$ ) are shown above/below the corresponding arrows. Only Aboriginal Australian individuals with low European ancestry were included in this analysis. In this model, we estimated parameters specific to the settlement of Australia and New Guinea (numerical values shown in black); keeping all the other demographic parameters set to the point estimates shown in Fig. 4 (numerical value shown in grey here). Only admixture events involving proportions  $>0.5\%$  are shown. The inferred parameters were scaled using a mutation rate of  $1.25 \times 10^{-8}$  per generation per site<sup>41</sup> and a generation time of 29 years corresponding to the average hunter-gatherer generation interval for males and females<sup>42</sup>. See Supplementary Information section S07 for further details.

observed SFS, suggesting that the ancestors of the sampled Aboriginal Australians diverged from a single ancestral population. This conclusion is also supported by MDS analyses (Fig. 2b), as well as by estimation of ancestry proportion<sup>58</sup> where all Aboriginal Australians form a cluster distinct from the Papuan populations (Extended Data Fig. 1, Supplementary Information section S05). Additionally, it is supported by outgroup  $f_3$  analyses, where all Aboriginal Australians are largely equidistant from Papuans when adjusting for recent admixture (Extended Data Fig. 2b). Thus, our results, based on 83 Pama-Nyungan speakers, do not support earlier claims of multiple ancestral migrations into Australia giving rise to contemporary Aboriginal Australian diversity<sup>24</sup>.

The SFS analysis indicates that there was a bottleneck in the ancestral Australo-Papuan population  $\sim 50$  kya (95% confidence intervals (CI) 35–54 kya, Supplementary Information section S07), which overlaps with archaeological evidence for the earliest occupation of both Sunda and Sahul 47–55 kya<sup>2–4</sup>. We further infer that the ancestors of Pama-Nyungan speakers and Highland Papuans diverged  $\sim 37$  kya (95% CI 25–40 kya, Fig. 3, Supplementary Information section S07), which is in close agreement with results of multiple sequentially Markovian coalescent (MSMC) analyses (Extended Data Fig. 2c, Supplementary Information section S08), a method estimating cross coalescence rates between pairs of populations based on individuals' haplotypes<sup>25</sup>. This result is also in agreement with previous estimates, for example, based on SNP array data<sup>18</sup>.

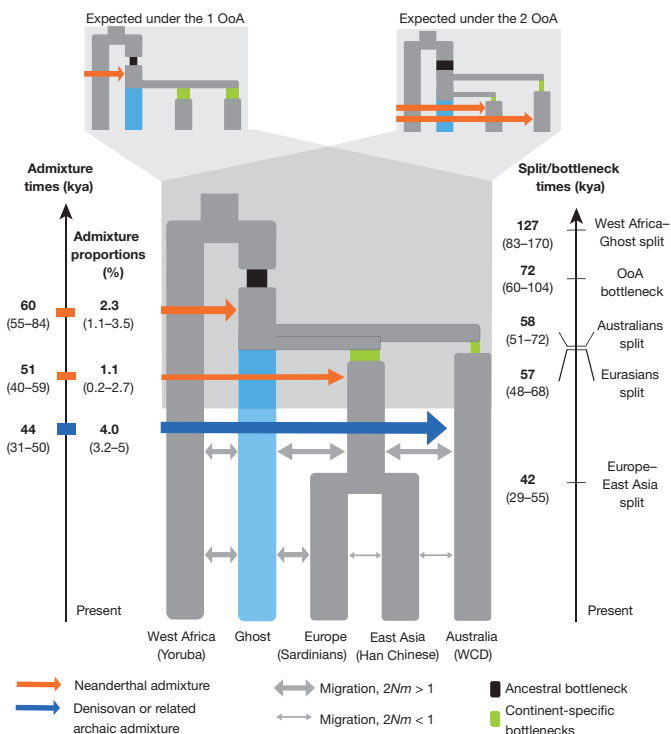
## Archaic admixture

We characterized the number, timing and intensity of archaic gene-flow events using three complementary approaches: SFS-based (Supplementary Information section S07), a goodness-of-fit analysis combining D-statistics (Supplementary Information section S09), and a method that infers putatively derived archaic 'haplotypes' (Supplementary Information section S10). Aboriginal Australian and Papuan genomes show an excess of putative Denisovan introgressed sites (Extended Data Fig. 3a, Supplementary Information section S11), as well as substantially more putative Denisovan-derived haplotypes (PDHs) than other non-Africans (Extended Data Fig. 3b, Supplementary Information section S10). The number and total length of those putative haplotypes vary considerably across samples. However, the estimated number of PDHs correlates almost perfectly ( $r^2 = 0.96$ ) with the estimated proportion of Australo-Papuan ancestry in each individual (Extended Data Fig. 3c). We found no significant difference in the distribution of the number of PDHs or the average length of PDHs between putatively unadmixed Aboriginal Australians and Papuans (Mann–Whitney  $U$ -test,  $P > 0.05$ ). Moreover, the genetic differentiation between WCD and Papuans was also similar for both autosomal SNPs and PDHs, with  $F_{ST}$  values around 0.12. Taken together, these analyses provide evidence for Denisovan admixture predating the population split between Aboriginal Australians and Papuans (see also refs 26, 53) and widespread recent Eurasian admixture in Aboriginal Australians (Fig. 2a, b, Supplementary Information section S05). By constraining Denisovan admixture as having occurred before the Aboriginal Australian–Papuan divergence, the SFS-based approach results in an admixture estimate of  $\sim 4.0\%$  (95% CI 3.3–5.0%, Fig. 4, Supplementary Information section S07), similar to that obtained by D-statistics ( $\sim 5\%$ , Supplementary Information section S09). The SFS analyses further suggest that Denisovan/Australo-Papuan admixture took place  $\sim 44$  kya (95% CI 31–50 kya, Supplementary Information section S07), a date that overlaps with an estimate from a more recent study<sup>54</sup>.

The SFS analysis also provides evidence for a primary Neanderthal admixture event ( $\sim 2.3\%$ , 95% CI 1.1–3.5%) taking place in the ancestral population of all non-Africans  $\sim 60$  kya (95% CI 55–84 kya, Fig. 4, Supplementary Information section S07). Although we cannot estimate absolute dates of archaic admixture from the lengths of PDHs and putative Neanderthal-derived haplotypes (PNHs) in our samples, we can obtain a relative date. We found that, for putatively unadmixed Aboriginal Australians and HGDP-Papuans, the average PNH and PDH lengths are 33.8 kb and 37.4 kb, respectively (Extended Data Fig. 3b). These are significantly different from each other ( $P = 9.65 \times 10^{-6}$  using a conservative sign test), and suggest that the time since Neanderthal admixture was about 11% greater than the time since Denisovan admixture, roughly in line with our SFS-based estimates for the Denisovan pulse (31–50 kya, Fig. 4) versus the primary pulse of Neanderthal admixture (55–84 kya). The SFS analysis also indicates that the main Neanderthal pulse was followed by a further 1.1% (95% CI 0.2–2.7%, Fig. 4, Supplementary Information section S07) pulse of Neanderthal gene flow into the ancestors of Eurasians. Finally, using our SFS- and haplotype-based approaches, we explored additional models involving complex structure among the archaic populations. We found suggestive evidence that the archaic contribution could be more complex than the model involving the discrete Denisovan and Neanderthal admixture pulses<sup>8,9</sup> shown in Fig. 4 (Supplementary Information sections S07, S10).

## Out of Africa

To investigate the relationship of Australo-Papuan ancestors with other world populations, we computed D-statistics<sup>22</sup> of the form ( $H1$  = Aboriginal Australian,  $H2$  = Eurasian),  $H3$  = African) and ( $H1$  = Aboriginal Australian,  $H2$  = Eurasian),  $H3$  = Ust'-Ishim). Several of these were significantly positive (Supplementary Information section S09), suggesting that Africans and Ust'-Ishim—the 45,000-year-old remains of a modern human from Asia<sup>27</sup>—are both closer to



**Figure 4 | Out of Africa.** We used a likelihood-based approach to investigate whether the joint SFS supports the one-wave (1 OoA) or two-wave (2 OoA) scenarios. The maximum likelihood estimates are indicative of which scenario is best supported. As shown on the top left inset, under the 1 OoA scenario we expect (i) the presence of an ancestral bottleneck (in black); (ii) a relatively large Neanderthal admixture pulse shared by the ancestors of all non-Africans; and (iii) overlapping divergence times of the ancestors of Aboriginal Australians and Eurasians. In contrast, the top right inset shows parameters expected under a 2 OoA scenario: (i) a limited/absent ancestral bottleneck (in black) in the ancestors of all non-Africans; (ii) no shared Neanderthal admixture in the ancestors of all non-Africans; (iii) distinct divergence times for Aboriginal Australians and Eurasians. The main population tree shows the best fitting topology, which supports the 1 OoA scenario, and maximum likelihood estimates (MLEs) for the divergence and admixture times and the admixture proportions (with 95% CI obtained by non-parametric block bootstrap shown within parentheses). We assume that the OoA event is associated with the ancestral bottleneck. The ‘Ghost’ population represents an unsampled population related to Yoruba that is the source of the out-of-Africa event(s). Our results suggest that these two African populations split significantly earlier (~125 kya) than the estimated time of dispersals into Eurasia. Note that under a 1 OoA scenario, this ghost population becomes, after the ancestral bottleneck, the ancestral population of all non-Africans that admixed with Neanderthals. Arrow thicknesses are proportional to the intensity of gene flow and the admixture proportions, and only admixture events involving proportions >0.5% are displayed. The inferred parameters were scaled as for Fig. 3. See Supplementary Information section S07 for further details.

Eurasians than to Aboriginal Australians. These findings are in agreement with a model of Eurasians and Australo-Papuan ancestors dispersing from Africa in two independent waves. However, when correcting for a moderate amount of Denisovan admixture, Aboriginal Australians and Eurasians become equally close to Ust'-Ishim, as expected in a single OoA scenario (Supplementary Information section S09). Similarly, the D-statistics for ((H1 = Aboriginal Australian, H2 = Eurasian), H3 = African) became much smaller after correcting for Denisovan admixture. Additionally, a goodness-of-fit approach combining D-statistics across worldwide populations indicates stronger support for two waves OoA, but when taking Denisovan admixture into account, a one-wave scenario fits the observed D-statistics equally well (Extended Data Fig. 4a, b, Supplementary Information section S09).

To investigate the timing and number of OoA events giving rise to present-day Australo-Papuans and Eurasians further, we used the observed SFS in a model-based composite likelihood framework<sup>23</sup>. When considering only modern human genomes, we find evidence for two waves OoA, with a dispersal of Australo-Papuans ~14,000 years before Eurasians (Supplementary Information section S07). However, when explicitly taking into account Neanderthal and Denisovan introgression into modern humans<sup>9,20</sup>, the SFS analysis supports a single origin for the OoA populations marked by a bottleneck ~72 kya (95% CI 60–104 kya, Fig. 4, Supplementary Information section S07). This scenario is reinforced by the observation that the ancestors of Australo-Papuans and Eurasians share a 2.3% (95% CI 1.1–3.5%) Neanderthal admixture pulse. Furthermore, modern humans have both a linkage disequilibrium decay rate and a number of predicted deleterious homozygous mutations (recessive genetic load) that correlate with distance from Africa (Supplementary Information sections S05, S11, Extended Data Fig. 5), again consistent with a single African origin.

The model best supported by the SFS analysis also suggests an early divergence of Australo-Papuans from the ancestors of all non-Africans, in agreement with two colonization waves across Asia<sup>8,18</sup>. Under our best model, Australo-Papuans began to diverge from Eurasians ~58 kya (95% CI 51–72 kya, Fig. 4, Supplementary Information section S07), whereas Europeans and East Asians diverged from each other ~42 kya (95% CI 29–55 kya, Fig. 4, Supplementary Information section S07), in agreement with previous estimates<sup>7,18,28</sup>. We find evidence for high levels of gene flow between the ancestors of Eurasians and Australo-Papuans, suggesting that, after the fragmentation of the OoA population (‘Ghost’ in Fig. 4) 57–58 kya, the groups remained in close geographical proximity for some time before Australo-Papuan ancestors dispersed eastwards. Furthermore, we find evidence for gene flow between sub-Saharan Africans and Western Eurasians after ~42 kya, in agreement with previous findings<sup>28</sup>.

MSMC analyses suggest that the Yoruba/Australo-Papuans and the Yoruba/Eurasians cross-coalescence rates are distinct, implying that the Yoruba and Eurasian gene trees across the genome have, on average, more recent common ancestors (Extended Data Fig. 4c, Supplementary Information section S08). We show through simulations that these differences cannot be explained by typical amounts of archaic admixture (<20%, Extended Data Fig. 4d). Moreover, the expected difference in phasing quality among genomes is not sufficient to explain this pattern fully (Supplementary Information section S08). While a similar separation in cross coalescence rate curves is obtained when comparing Eurasians and Australo-Papuans with Dinka, we find that, when comparing Australo-Papuans and Eurasians with San, the cross coalescence curves overlap (Extended Data Fig. 4c). We also find that the inferred changes in effective population size through time of Aboriginal Australians, Papuans, and East Asians are very similar until around 50 kya, including a deep bottleneck around 60 kya (Extended Data Fig. 6a). Taken together, these MSMC results are consistent with a split of both Australo-Papuans and Eurasians from a single African ancestral population, combined with gene flow between the ancestors of Yoruba or Dinka (but not San) and the ancestors of Eurasians that is not shared with Australo-Papuans. These results are qualitatively in line with the SFS-based analyses (see Fig. 4). While our results do not exclude the possibility of an earlier OoA expansion, they do indicate that any such event left little trace in the genomes of modern Australo-Papuans, in line with conclusions from related work appearing alongside this study<sup>55,56</sup>.

### Genetic structure of Aboriginal Australians

Uniparental haplogroup diversity in this dataset (Extended Data Table 1, Supplementary Information section S12) is consistent with previous studies of mitochondrial DNA (mtDNA) and Y chromosome variation in Australia and Oceania<sup>29</sup>, including the presence of typically European, Southeast and East Asian lineages<sup>30</sup>. The combined results



provide important insights into the social structure of Aboriginal Australian societies. Aboriginal Australians exhibit greater between-group variation for mtDNA (16.8%) than for the Y chromosome (11.3%), in contrast to the pattern for most human populations<sup>31</sup>. This result suggests higher levels of male- than female-mediated migration, and may reflect the complex marriage and post-marital residence patterns among Pama–Nyungan Australian groups<sup>32</sup>. As expected (Supplementary Information section S02), the inferred European ancestry for the Y chromosome is much greater than that for mtDNA (31.8% versus 2.4%), reflecting male-biased European gene flow into Aboriginal Australian groups during the colonial era.

On an autosomal level, we find that genetic relationships within Australia reflect geography, with a significant correlation ( $r_{\text{GEN,GEO}} = 0.77$ ,  $P < 0.0005$ , Extended Data Fig. 7b) between the first two dimensions of an MDS analysis on masked genomes and geographical location (Supplementary Information section S13). Populations from the centre of the continent occupy genetically intermediate positions (Extended Data Fig. 7a, b). A similar result is observed with an  $F_{\text{ST}}$ -based tree for the masked genomic data (Extended Data Fig. 7c, Supplementary Information section S05) as well as in analyses of genetic affinity based on  $f_3$  statistics (Extended Data Fig. 2a), suggesting a population division between northeastern and southwestern groups. This structure is further supported by SFS analyses showing that populations from southwestern desert and northeastern regions diverged as early as  $\sim 31$  kya (95% CI 10–32 kya, Fig. 3), followed by limited gene flow (estimated scaled migration rate ( $2Nm$ )  $\sim 0.01$ , 95% CI 0.00–11.25). An analysis of the major routes of gene flow within the continent supports a model in which the Australian interior acted as a barrier to migration. Using a model inspired by principles of electrical engineering where gene flow is represented as a current flowing through the Australian continent and using observed  $F_{\text{ST}}$  values as a proxy for resistance, we infer that gene flow occurred preferentially along the coasts of Australia (Extended Data Fig. 7e–g, Supplementary Information section S13). These findings are consistent with a model of expansion followed by population fragmentation when the extreme aridity in the interior of Australia formed barriers to population movements during the LGM<sup>33</sup>.

We used MSMC on autosomal data and mtDNA Bayesian skyline plots<sup>34</sup> (BSP) to estimate changes in effective population size within Australia. The MSMC analyses provide evidence of a population expansion starting  $\sim 10$  kya in the northeast, while both MSMC and BSP indicate a bottleneck in the southwestern desert populations taking place during the past  $\sim 10,000$  years (Extended Data Fig. 6, Supplementary Information sections S08, S12). This is consistent with archaeological evidence for a population expansion associated with significant changes in socio-economic and subsistence strategies in Holocene Australia<sup>35</sup>.

European admixture almost certainly had not occurred before the late 18th century, but earlier East Asian and/or New Guinean gene flow into Australia could have taken place. We characterized the mode and tempo of gene flow into Aboriginal Australians using three different approaches (Supplementary Information sections S06, S07, S14). We used approximate Bayesian computation (ABC) to compare the observed mean and variance in the proportion of European, East Asian and Papuan admixture among Aboriginal Australian individuals to that computed from simulated datasets under various models of gene flow. We estimated European and East Asian admixture to have occurred approximately ten generations ago (Supplementary Information section S14), consistent with historical and ethnographic records. Consistent with this, a local ancestry approach suggests that European and East Asian admixture is more recent than Papuan admixture (Extended Data Fig. 8, Supplementary Information section S06). In addition, both ABC and SFS analyses indicate that the best-fitting model for the Aboriginal Australian–Papuan data is one of continuous but modest gene flow, mostly unidirectional from Papuans to Aboriginal Australians, and geographically restricted to northeast Aboriginal

Australians ( $2Nm = 0.41$ , 95% CI 0.00–20.35, Fig. 3, Supplementary Information section S07).

To investigate gene flow from New Guinea further, we conducted analyses on the Papuan ancestry tracts obtained from the local ancestry analysis. We inferred local ancestry as the result of admixture between four components: European, East Asian, Papuan and Aboriginal Australian (Supplementary Information section S06). The Papuan tract length distribution shows a clear geographic pattern (Extended Data Fig. 8b); we find a significant correlation of Papuan tract length variance with distance from WCD to other Aboriginal Australian groups ( $r = 0.64$ ,  $P < 0.0001$ ). The prevalence of short ancestry tracts of Papuan origin, compared to longer tracts of East Asian and European origin, suggests that a large fraction of the Papuan gene flow is much older than that from Europe and Asia, consistent with the ABC analysis (Supplementary Information section S14). We also investigated possible South Asian (Indian-related) gene flow into Aboriginal Australians, as reported recently<sup>18</sup>. However, we found no evidence of a component that can be uniquely assigned to Indian populations in the Aboriginal Australian gene pool using either admixture analyses or  $f_3$  and D-statistics (Supplementary Information section S05), even when including the original Aboriginal Australian genotype data from Arnhem Land. The different size and nature of the comparative datasets may account for this discrepancy.

### Pama–Nyungan languages and genetic structure

To investigate whether linguistic relationships reflect genetic relationships among Aboriginal Australian populations, we inferred a Bayesian phylogenetic tree for the 28 different Pama–Nyungan languages represented in this sample<sup>13</sup> (Extended Data Table 1, Supplementary Information section S15). The resulting linguistic and  $F_{\text{ST}}$ -based genetic trees (Extended Data Fig. 7c, d) share several well-supported partitions. For example, both trees indicate that the northeastern (CAI and WPA) and southwestern groups (ENY, NGA, WCD and WON) form two distinct clusters, while PIL, BDV and RIV are intermediate. A distance matrix between pairs of languages, computed from the language-based tree, is significantly correlated with geographic distances ( $r_{\text{GEO,LAN}} = 0.83$ , Mantel test two-tail  $P$  on 9,999 permutations = 0.0001, Supplementary Information section S13). This suggests that differentiation among Pama–Nyungan languages in Australia follows geographic patterns, as observed in other language families elsewhere in the world<sup>36</sup>. Furthermore, we find a correlation between linguistics and genetics ( $r_{\text{GEN,LAN}} = 0.43$ , Mantel test  $P < 0.0005$ , Supplementary Information section S13) that remains significant when controlling for geography ( $r_{\text{GEN,LAN,GEO}} = 0.26$ , partial Mantel test  $P < 0.0005$ , Supplementary Information section S13). This is consistent with language differentiation after populations lost (genetic) contact with one another. The correlation between the linguistic and genetic trees is all the more notable given the difference in time scales: the Pama–Nyungan family is generally accepted to have diversified within the last 6,000 years<sup>37</sup>, while the genetic estimates are two to five times that age. The linguistic tree thus cannot simply reflect initial population dispersals, but rather reflects a genetic structure that has a complex history, with initial differentiation 10–32 kya, localized population expansions (northeast) and bottlenecks (southwest)  $\sim 10$  kya, and subsequent limited gene flow from the northeast to the southwest. The latter may be the genetic signature that tracks the divergence of the Pama–Nyungan language family.

### Selection in Aboriginal Australians

To identify selection signatures specific to Aboriginal Australians, we used two different methods based on the identification of SNPs with high allele-frequency differences between Aboriginal Australians and other groups, similar to the population-branch statistics<sup>38</sup> (PBS, Supplementary Information section S16). First, we scanned the Aboriginal Australian genomes for loci with unusually large changes in allele frequency since divergence from Papuans, taking recent



admixture with Europeans and Asians into account ('global scan'). Second, we identified genomic regions showing high differentiation associated with different ecological regions within Australia ('local scan'; Supplementary Information section S16). Among the top ranked peaks (Extended Data Table 2) we found genes associated with the thyroid system (*NETO1*, seventh peak in the global scan, and *KCNJ2*, first peak in the local scan) and serum urate levels (eighth peak in the global scan). Thyroid hormone levels are associated with Aboriginal-Australian-specific adaptations to desert cold<sup>39</sup> and elevated serum urate levels with dehydration<sup>40</sup>. These genes are therefore candidates for potential adaptation to life in the desert. However, further studies are needed to associate putative selected genetic variants with specific phenotypic adaptations in Aboriginal Australians.

## Discussion

Australia has one of the longest histories of continuous human occupation outside Africa, raising questions of origins, relatedness to other populations, differentiation and adaptation. Our large-scale genomic data and analyses provide some answers but also raise new questions. We find that Aboriginal Australians and Eurasians share genomic signatures of an OoA dispersal—a common African ancestor, a bottleneck and a primary pulse of Neanderthal admixture. However, Aboriginal Australian population history diverged from that of other Eurasians shortly after the OoA event, and included private admixture with another archaic hominin.

Our genetic-based time estimates are relative, and to obtain absolute dates we relied on two rescaling parameters: the human mutation rate and generation time (assumed to be  $1.25 \times 10^{-8}$  per generation per site and 29 years, respectively, based on recent estimates<sup>41,42</sup>). Although the absolute estimates we report would need to be revised if these parameters were to change, the current values can be the starting point of future research and should be contextualized.

We find a relatively old divergence between the ancestors of Pama-Nyungan speakers and Highland Papuans, only ~10% younger than the European–East Asian split time. With the assumed rescaling parameters this corresponds to ~37 kya (95% CI 25–40 kya), implying that the divergence between sampled Papuans and Aboriginal Australians is older than the disappearance of the land bridge between New Guinea and Australia ~7–14.5 kya, and thus suggests ancient genetic structure in Sahul. Such structure may be related to palaeo-environmental changes leading up to the LGM. Sedimentary studies show that the large Lake Carpentaria (500 × 250 km, Fig. 1) formed ~40 kya, when sea levels fell below the 53-m-deep Arafura Sill<sup>43</sup>. Although Australia and New Guinea remained connected until the early Holocene, the flooding of the Carpentaria basin and its increasing salinity<sup>43</sup> may have thus promoted population isolation.

Our results imply that Aboriginal Australian groups are the descendants of the ancestral population that first colonized Australia<sup>8,44</sup>. They also indicate that the population that diverged from Papuans ~37 kya was ancestral to all Aboriginal Australian groups sampled in this study; yet, archaeological evidence shows that by 40–45 kya, humans were widespread within Australia (Fig. 1). Three non-exclusive scenarios could account for this observation: (1) the Aboriginal Australian ancestral population was widespread before the divergence from Papuans, maintaining gene flow across the continent; (2) it was deeply structured, and only one group survived to give rise to modern Aboriginal Australians; and (3) other groups survived, but the descendants are not represented in our sample. Additional genomes, especially from Tasmania and the non-Pama-Nyungan regions of the Northern Territory and Kimberley, as well as ancient genomes pre-dating European contact in Australia and other expansions across Southeast Asia<sup>17</sup>, may help resolve these questions in the future.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 October 2015; accepted 4 May 2016.

Published online 21 September 2016.

- Davidson, I. The colonization of Australia and its adjacent islands and the evolution of modern cognition. *Curr. Anthropol.* **51**, S177–S189 (2010).
- Clarkson, C. *et al.* The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. *J. Hum. Evol.* **83**, 46–64 (2015).
- O'Connell, J. F. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J. Archaeol. Sci.* **56**, 73–84 (2015).
- Barker, G. *et al.* The 'human revolution' in lowland tropical Southeast Asia: the antiquity and behaviour of anatomically modern humans at Niah Cave (Sarawak, Borneo). *J. Hum. Evol.* **52**, 243–261 (2007).
- Lahr, M. M. & Foley, R. Multiple dispersals and modern human origins. *Evol. Anthropol. Issues News Rev.* **3**, 48–60 (1994).
- Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc. Natl Acad. Sci. USA* **111**, 7248–7253 (2014).
- Wollstein, A. *et al.* Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20**, 1983–1992 (2010).
- Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334**, 94–98 (2011).
- Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
- Reeves, J. M. *et al.* Climate variability over the last 35,000 years recorded in marine and terrestrial archives in the Australian region: an OZ-INTIMATE compilation. *Quat. Sci. Rev.* **74**, 21–34 (2013).
- Hiscock, P. & Wallis, L. A. in *Desert Peoples* (eds Veth, P., Smith, M. & Hiscock, P.) 34–57 (Blackwell Publishing Ltd, 2005).
- Birdsell, J. B. *Microevolutionary Patterns in Aboriginal Australia: A Gradient Analysis of Clines*. (Oxford University Press, 1993).
- Bowern, C. & Atkinson, Q. Computational phylogenetics and the internal structure of Pama-Nyungan. *Language* **88**, 817–845 (2012).
- Dixon, R. M. W. *Australian Languages: Their Nature and Development*. (Cambridge University Press, 2002).
- Evans, N. & McConvell, P. in *Archaeology and Language II: Archaeological Data and Linguistic Hypotheses* (eds Blench, R. & Spriggs, M.) Ch. 7 (Routledge, 1999).
- Hiscock, P. *Archaeology of ancient Australia*. (Routledge, 2008).
- Bellwood, P. *First Migrants: Ancient Migration in Global Perspective*. (Wiley-Blackwell, 2013).
- Pugach, I., Delfin, F., Gunnarsdóttir, E., Kayser, M. & Stoneking, M. Genome-wide data substantiate Holocene gene flow from India to Australia. *Proc. Natl Acad. Sci. USA* **110**, 1803–1808 (2013).
- Ellinghaus, K. Absorbing the 'Aboriginal problem': controlling interracial marriage in Australia in the late 19th and early 20th centuries. *Aborig. Hist.* **27**, 183–207 (2003).
- Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
- Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973–983 (2014).
- Patterson, N. J. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
- Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. Robust demographic inference from genomic and SNP data. *PLoS Genet.* **9**, e1003905 (2013).
- Thorne, A. G. in *The Origin of the Australians* (eds Kirk, R. L. & Thorne, A. G.) 95–112 (Canberra: Australian Institute of Aboriginal Studies, 1976).
- Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
- Qin, P. & Stoneking, M. Denisovan Ancestry in East Eurasian and Native American Populations. *Mol. Biol. Evol.* **32**, 2665–2674 (2015).
- Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
- Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695 (2009).
- Bergström, A. *et al.* Deep roots for Aboriginal Australian Y chromosomes. *Curr. Biol.* **26**, 809–813 (2016).
- Hudjashov, G. *et al.* Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc. Natl Acad. Sci. USA* **104**, 8726–8730 (2007).
- Lippold, S. *et al.* Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investig. Genet.* **5**, 13 (2014).
- Radcliffe-Brown, A. R. The social organization of Australian tribes. *Oceania* **1**, 34–63 (1930).
- Veth, P. Islands in the interior: a model for the colonization of Australia's arid zone. *Archaeol. Ocean.* **24**, 81–92 (1989).
- Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–1192 (2005).

35. Lourandos, H. & David, B. in *Bridging Wallace's Line: the Environmental and Cultural History and Dynamics of the SE Asian-Australasian Region* (eds Kershaw, A. P., David, B., Tapper, N., Penny, D. & Brown, J.) *Advances in GeoEcology* **34**, 97–118 (2002).
36. Cavalli-Sforza, L. L. Genes, peoples and languages. *Proc. Natl Acad. Sci. USA* **91**, 7719–7724 (1997).
37. Evans, N. & Jones, R. in *Archaeology and linguistics: Aboriginal Australia in global perspective* (Oxford University Press Australia, 1997).
38. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
39. Qi, X., Chan, W. L., Read, R. J., Zhou, A. & Carrell, R. W. Temperature-responsive release of thyroxine and its environmental adaptation in Australians. *Proc. Biol. Sci.* **281**, 20132747 (2014).
40. Tin, A. *et al.* Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel UAT1 loss-of-function allele. *Hum. Mol. Genet.* **20**, 4056–4068 (2011).
41. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
42. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
43. Holt, S. *Palaeoenvironments of the Gulf of Carpentaria from the Last Glacial Maximum to the Present, as Determined by Foraminiferal Assemblages*. PhD thesis, Univ. Wollongong (2005).
44. Heupink, T. H. *et al.* Ancient mtDNA sequences from the First Australians revisited. *Proc. Natl Acad. Sci. USA* **113**, 6892–6897 (2016).
45. Horton, D. (ed) *The Encyclopaedia of Aboriginal Australia*. (Aboriginal Studies Press, 1994).
46. Miglino, A. B. *et al.* Evolution of the pygmy phenotype: evidence of positive selection from genome-wide scans in African, Asian, and Melanesian pygmies. *Hum. Biol.* **85**, 251–284 (2013).
47. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
48. Wall, J. D. *et al.* Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209 (2013).
49. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
50. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
51. Wang, C. *et al.* Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat. Appl. Genet. Mol. Biol.* **9**, 13 (2010).
52. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
53. Skoglund, P. & Jakobsson, M. Ancient human ancestry in East Asia. *Proc. Natl. Acad. Sci. USA* **108**, 18301–18306 (2011).
54. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr Biol.* **26**, 1241–1247 (2016).
55. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* <http://dx.doi.org/10.1038/nature18964> (this issue).
56. Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* <http://dx.doi.org/10.1038/nature19792> (this issue).
57. Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516–528 (2011).
58. Cheng, J. Y., Mailund, T., & Nielsen, R. Ohana, a tool set for population genetic analyses of admixture components. *bioRxiv* doi:10.1101/071233 (2016).

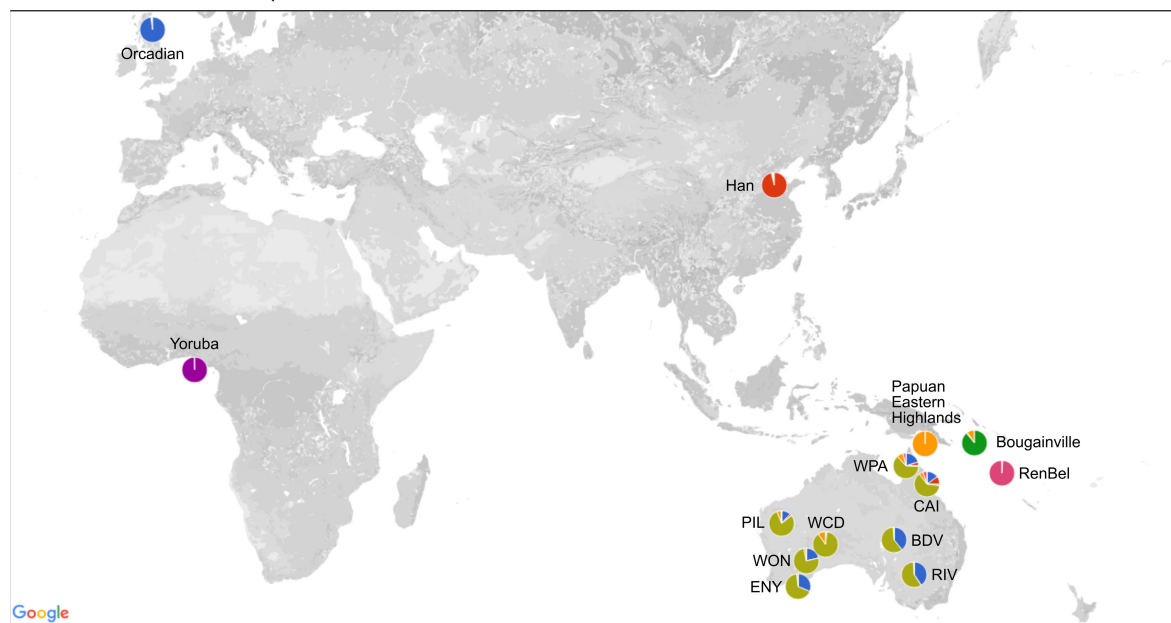
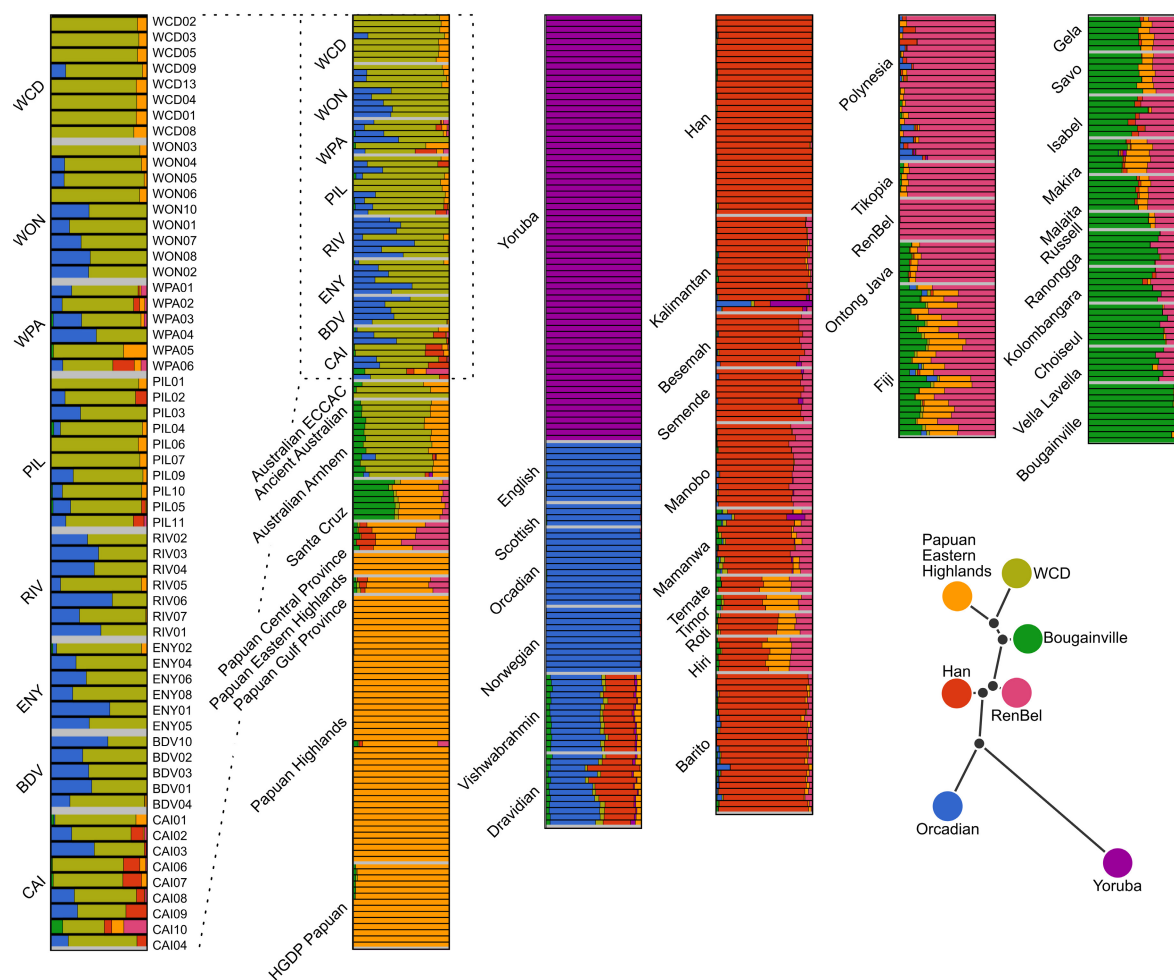
**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank all sample donors for contributing to this study. We thank Macrogen (<http://www.macrogen.com/>) for sequencing of the Aboriginal Australian genomes, M. Rasmussen, C. Der Sarkissian, M. Allentoft, D. Cooper, R. Gray, S. Greenhill, A. Seguin-Orlando, T. Carstensen, M. Przeworski, J. D. Jensen and L. Orlando for helpful discussions. We thank E. Thorsby for sample collection and contributing the DNA extract for the P2077 genome, I. Lissimore for support with data storage and distribution. We thank T. Parks, K. Auckland, K. Robson, A. V. Hill, J. B. Clegg, D. Higgs, D. J. Weatherall and M. Alpers for assistance in sample collection and discussion. L.E., V.C.S., I.A., I.D. and S.P. are grateful to the High Performance Computation platform of the University of Bern for providing access to the UBELIX cluster. This work was supported by the Danish National Research Foundation, the Lundbeck Foundation, the KU2016 grant and the Australian Research Council. A.-S.M. was supported by an ambition grant with reference PZ00P3\_154717 from the Swiss National Science Foundation (SNSF). M.C.W. was supported by the Australian Research Council (ARC) Discovery grants DP110102635 and DP140101405 and by a Linkage grant LP140100387. V.C.S., I.D. and S.P. were supported by SNSF grants to L.E. with references 31003A-143393 and CRSI13\_141940. O.L. was supported by a Ramón y Cajal grant from the

Spanish Ministerio de Economía y Competitividad (MINECO) with reference RYC-2013-14797 and by a BFU2015-68759-P (MINECO/FEDER) grant. I.A. was supported by a grant with reference SFRH/BD/73150/2010 from the Portuguese Foundation for Science and Technology (FCT). A.B., S.Sc., Y.X., C.T.-S. and R.D. were supported by a Wellcome Trust grant with reference WT098051. E.M., C.Ba., I.P., S.N. and M.St. acknowledge the Max Planck Society. S.Su. was supported by an ARC Discovery grant with reference DP140101405. J.L.W. was supported by a PhD scholarship from Griffith University. A.A. acknowledges the Villum foundation. I.M. was supported by a grant from the Danish Council for Independent Research with reference DFF-4090-00244. J.V.M.-M. acknowledges the Consejo Nacional de Ciencia y Tecnología (Mexico) for funding. N.B. and F.-X.R. were supported by the French Ministry of Foreign and European Affairs and French ANR with the grant ANR14-CE31-0013-01. S.B. was supported by a Novo Nordisk Foundation grant with reference NNF14CC0001. P.G. and A.B.M. were supported by a Leverhulme Programme grant number RP2011-R-045 to A.B.M. at UCL Department of Anthropology and M.G.T. at UCL Department of Genetics, Evolution and Environment. A.J.M. was supported by a Wellcome Trust grant with reference 106289/Z/14/Z. M.M. acknowledges the EU European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre; Estonian Institutional Research grant IUT24-1. M.G.T. was supported by a Wellcome Trust Senior Investigator Award with grant number 100719/Z/12/Z. S.J.O. was supported by a Wellcome Trust Core Award Grant Number 090532/Z/09/Z. A.Man. was supported by an ERC Consolidator Grant 647787 'LocalAdaptation'. M.E.P. would like to acknowledge the cardio-metabolic research cluster at Jeffrey Cheah School of Medicine & Health Sciences, Monash University Malaysia and Ministry of Science, Technology & Innovation, Malaysia for research grant 100-RM1/BIOTEK 16/6/2B. M.H.S. was supported by a grant from the Danish Independence Research Council with reference FNU 12-125062. R.A.F. was supported by the Leverhulme Trust. M.M.L. is supported by an ERC Advanced Grant 295907 'In-Africa'. C.Bo. was supported by USA National Science Foundation (NSF) grants BCS-0844550 and BCS-1423711, awarded to C.Bo. and Yale University. T.M. was supported by a grant from the Danish Independence Research Council with reference FNU 1323-00749. M.S.S. was supported by a Wellcome Trust grant with reference WT098051. L.E. was supported by Swiss NSF grant number 31003A-143393, D.M.L. was supported by ARC Discovery Grants DP110102635 and DP140101405 and Linkage grants LP140100387, LP120200144 and LP150100583. E.W. is grateful to St John's College in Cambridge for help and support.

**Author Contributions** G.A., J.Y.C., J.E.C., T.H.H., E.M., S.P., S.R., S.Sc., S.Su. and J.L.W. contributed equally and are listed alphabetically in the author list; A.A., C.Ba., I.D., A.E., A.Mar., I.M. and I.P. contributed equally and are listed alphabetically in the author list; T.S.K., I.P.L., J.V.M.-M., S.N., F.R., M.Si. and Y.X. contributed equally and are listed alphabetically in the author list. E.W. and D.M.L. initially conceived and headed the project. L.E. led the genetic load and the SFS-based demographic analyses. M.S.S. headed the research at the Wellcome Trust Sanger Institute. A.-S.M. planned and coordinated the genetic analyses and the sequencing of the Aboriginal Australian genomes. C.M., J.L.W., T.H.H., P.F.C., W.C., G.F., D.I., B.L., A.L., P.J.M., L.M., D.R., T.W., C.W., J.D., M.C.W. and E.W. collaborated with local groups to collect Aboriginal Australian samples. N.B., P.G., G.K., M.L., A.J.M., A.B.M., W.P., F.-X.R., P.S., M.G.T. and S.J.O. collaborated with local groups to collect Papuan samples. S.E. collaborated with local groups to collect the Rapanui sample. A.Mar. extracted DNA for the Aboriginal Australian genomes. M.S.S., A.B. and C.T.-S. coordinated the design and sequencing of the Papuan genomes. O.L., V.C.S., I.A., A.-S.M., A.B., G.A., J.Y.C., J.E.C., T.H.H., E.M., S.P., S.R., S.Sc., S.Su., J.L.W., A.A., C.Ba., I.D., A.E., A.Man., I.M., I.P., T.S.K., I.P.L., J.V.M.-M., S.N., F.R., M.Si., F.A., S.B., L.E., J.D.W. and T.M. analysed genetic data. C.Bo. collected and analysed linguistic data. L.E., E.W., D.M.L., Y.X., M.E.P., C.T.-S., R.D., M.S.S., A.Man., M.H.S., T.M., M.St. and R.N. supervised genetic analyses. M.C.W., C.M., W.C., G.F., D.I., B.L., A.L., P.J.M., L.M., D.R., T.W., C.W., E.A.M.-S., M.M., M.E.P., S.J.O., J.D., A.B.M., R.A.F. and M.M.L. provided archaeological, anthropological and historical context. A.-S.M., V.C.S., O.L., I.A., A.B., M.M.L., R.N., L.E., D.M.L. and E.W. wrote the manuscript with critical input from G.A., T.H.H., E.M., S.Sc., S.Su., J.L.W., C.Ba., A.E., I.P., E.A.M.-S., M.S.S., S.J.O., C.T.-S., R.D., M.G.T., J.D., A.Man., M.H.S., R.A.F., C.Bo., J.D.W., T.M., M.St. and all other coauthors. A.-S.M., V.C.S., O.L., I.A. and A.B. revised and compiled the Supplementary Information.

**Author Information** The Aboriginal Australian and Papuan whole genome sequence data generated in this study have been deposited at the European Genome-phenome Archive (EGA, <http://www.ebi.ac.uk/ega/>), which is hosted by the EBI, under the accession numbers EGAS00001001766 and EGAS00001001247, respectively. The Papuan SNP array data generated in this study can be found under <http://geogenetics.ku.dk/latest-news/allelyheader/2016/data>. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.W. ([ewillerslev@snm.ku.dk](mailto:ewillerslev@snm.ku.dk)), D.M.L. ([d.lambert@griffith.edu.au](mailto:d.lambert@griffith.edu.au)), L.E. ([laurent.excoffier@iee.unibe.ch](mailto:laurent.excoffier@iee.unibe.ch)) and M.S.S. ([ms23@sanger.ac.uk](mailto:ms23@sanger.ac.uk)).



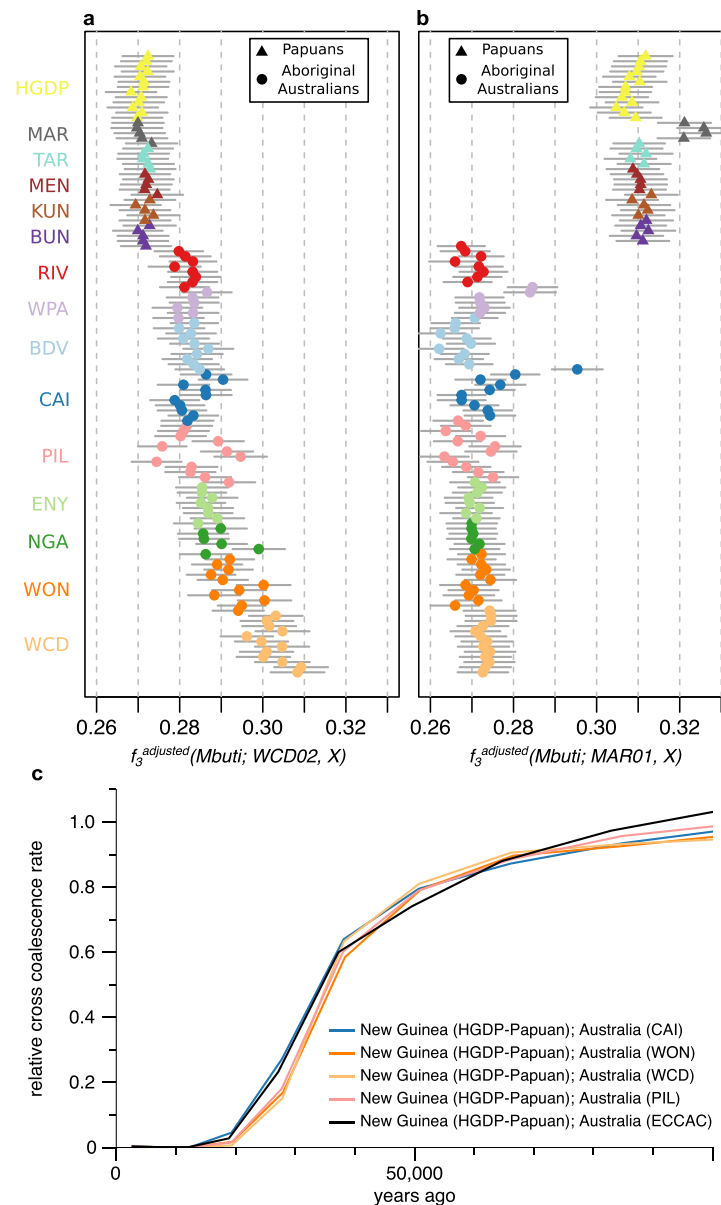
Map data © 2016 Google. INEGI · Phylogenetic trees: <http://jade-cheng.com/trees/>

Extended Data Figure 1 | See next page for caption.



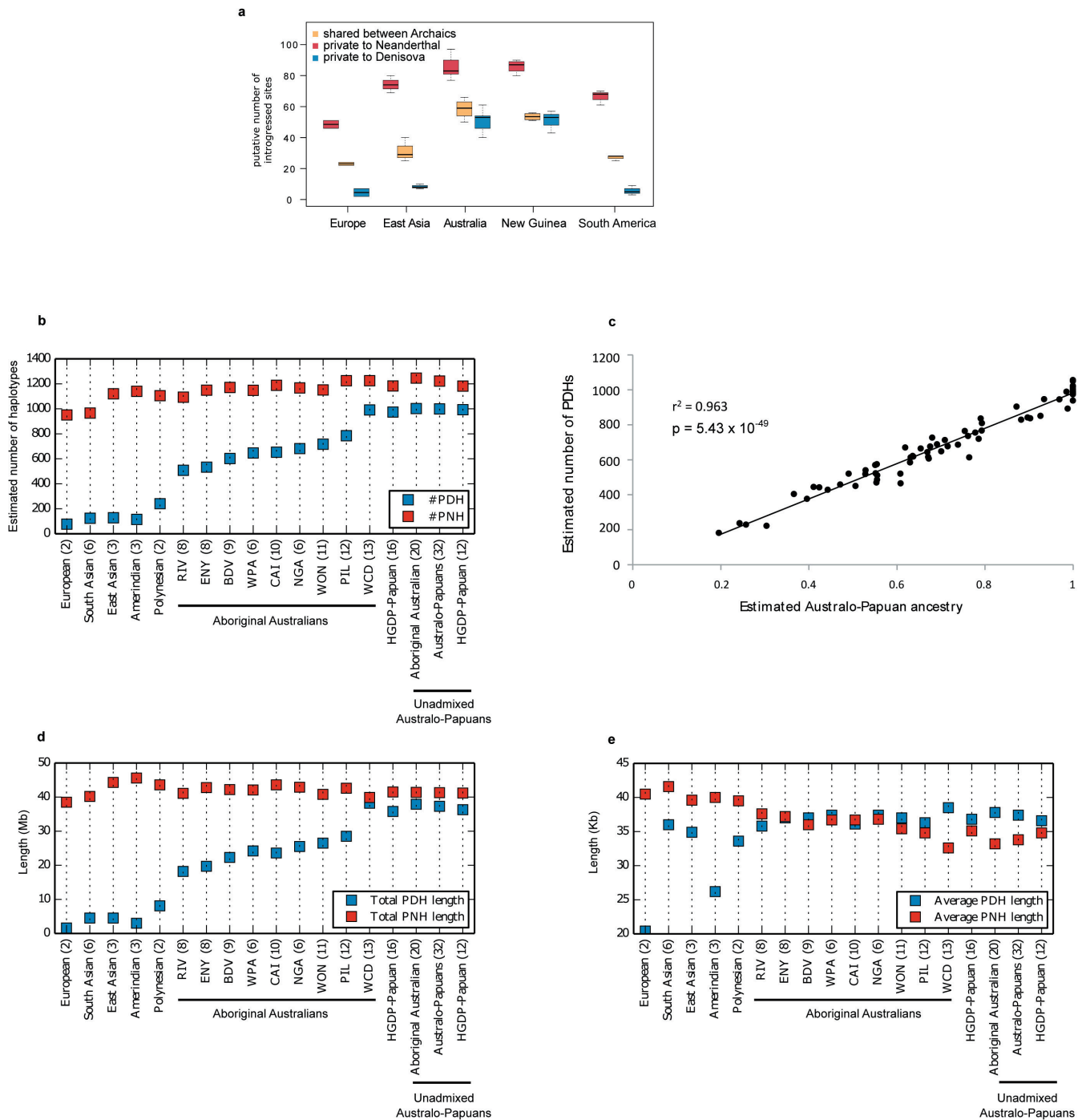
**Extended Data Figure 1 | Per-individual admixture proportions of  $K=7$  ancestral components including Aboriginal Australians, New Guineans, Europeans, Africans, Melanesians and Polynesians.** The genome of each individual is depicted as a bar and is coloured according to the estimated genome-wide proportions of ancestry components. An unrooted tree showing the relationships between the identified ancestral components is also estimated by our method. Each ancestry has been labelled with the name of the population (see also map) showing the highest fraction of that ancestral component. The cross-validation

error is minimized for this value of  $K$  for fivefold cross-validation. The rooted tree supports the shared genetic origin of Aboriginal Australians, Papuans and Bougainvilleans. Note that only individuals with more than 50% of Aboriginal Australian ancestry in their genomes (defined in Supplementary Information section S06) were included in the analyses. Refer to ref. 58 and Supplementary Information section S05 for details about the method and the analysis. Map data ©2016 Google, INEGI. Tree constructed with <http://jade-cheng.com/trees/>.



**Extended Data Figure 2 | Genetic relationships of Aboriginal Australians and Papuans.** **a**, Genetic affinities between a western central desert (WCD02) genome and Aboriginal Australians and Papuans. Outgroup  $f_3$  statistics between WCD02 and all other Aboriginal Australians and Highland Papuan individuals that were whole-genome sequenced for this study, using the genotypes called from the sequencing data. Because the widespread recent admixture in Aboriginal Australians has large confounding effects on the  $f_3$  statistics, the values were adjusted using the slope coefficient from a simple linear regression model fitted to the relationship between  $f_3$  and the fraction of non-indigenous (that is, neither Aboriginal Australian nor Papuan) ancestry in each individual genome. The adjusted  $f_3$  statistics display a genetic gradient that separates western and eastern Aboriginal Australian populations. However, we find no differences between Papuan population samples in their level of Aboriginal Australian affinity (Kruskal–Wallis test,  $P = 0.083$ ). Horizontal lines correspond to  $\pm 1$  standard error. **b**, Genetic affinities between a Papuan highlander genome and Aboriginal Australians and Papuans. The Papuan highlander sample MAR01 from the Marawaka area was arbitrarily chosen as a reference point for this analysis.  $f_3$  values were adjusted for recent admixture as in **a**. All Aboriginal Australian groups display a similar level of Highland Papuan affinity (with the exception

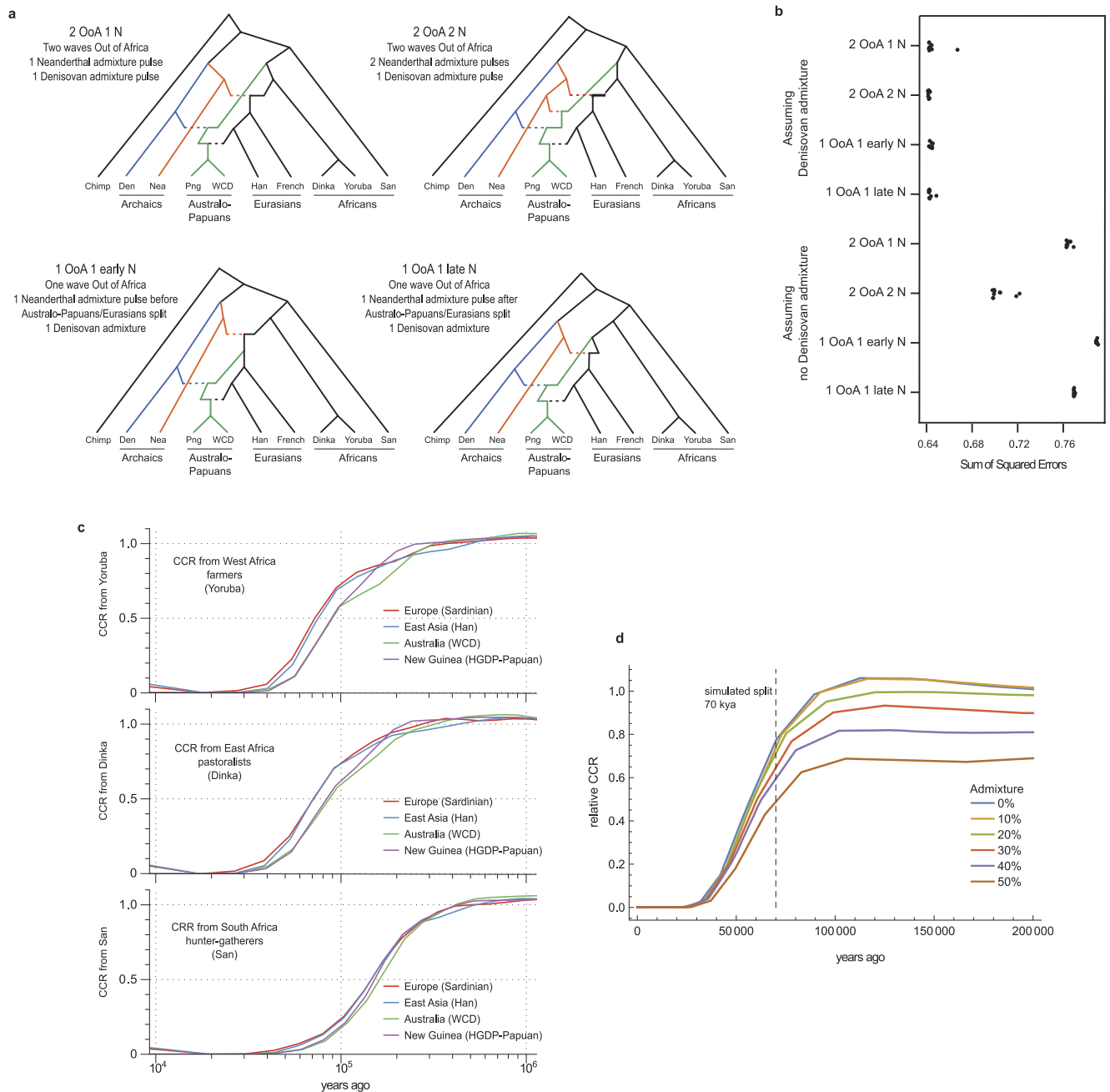
of three outlier individuals from the north-eastern WPA and CAI populations: WPA06, WPA05 and CAI10, the latter two of which are known to have at least one parent with origins in Papua New Guinea or the Torres Strait Islands). While some differences between groups are actually statistically significant (Kruskal–Wallis test,  $P = 0.0002$ , after removing the three outliers), which could be consistent with, for example, low levels of Papuan gene flow into some Aboriginal Australian groups (see Supplementary Information sections S06 and S07), we caution that some of these differences are probably due to imperfect adjustment for Eurasian admixture (the adjusted  $f_3$  is highest in the WCD population, which has the least Eurasian admixture). Horizontal lines correspond to  $\pm 1$  standard error. **c**, MSMC analyses. Linear interpolation through the midpoints of the time intervals of the relative cross coalescence rate estimates from MSMC<sup>25</sup> using pairs of individuals including one HGDP-Papuan and one other individual as indicated. We used CAI01, PIL06, WCD01, WON03 and an ECCAC sample for this analysis (see Supplementary Information section S08 for details). The MSMC results were scaled using a mutation rate of  $1.25 \times 10^{-8}$  per generation per site as suggested in ref. 41 and a generation time of 29 years, corresponding to the average hunter–gatherer generation interval for males and females<sup>42</sup>.



**Extended Data Figure 3 | Introgressed archaic sites and putative Denisovan and Neanderthal haplotypes.** **a**, Distribution of number of putative introgressed sites per individual from archaic humans. The number of Neanderthal-specific introgressed sites per individual increases from Europe to Australia, and then decreases in Amerindians, which is consistent with recurrent Neanderthal (or Neanderthal-related archaic) gene flow during the expansion into Eurasia. Our results are thus indicative of several pulses of Neanderthal gene flow into modern humans, as inferred previously<sup>48–50</sup>. We note, however, that the apparent high levels of Neanderthal-specific introgressed sites in Australo-Papuans can be explained by the expected number of misclassified Neanderthal introgressed sites resulting from the shared ancestry with Denisovans (see Supplementary Information section S11 for details). **b–e**, Putative Denisovan (PDH) and Neanderthal haplotypes (PNH). The putative haplotypes correspond to clusters (four or more SNPs spanning at least

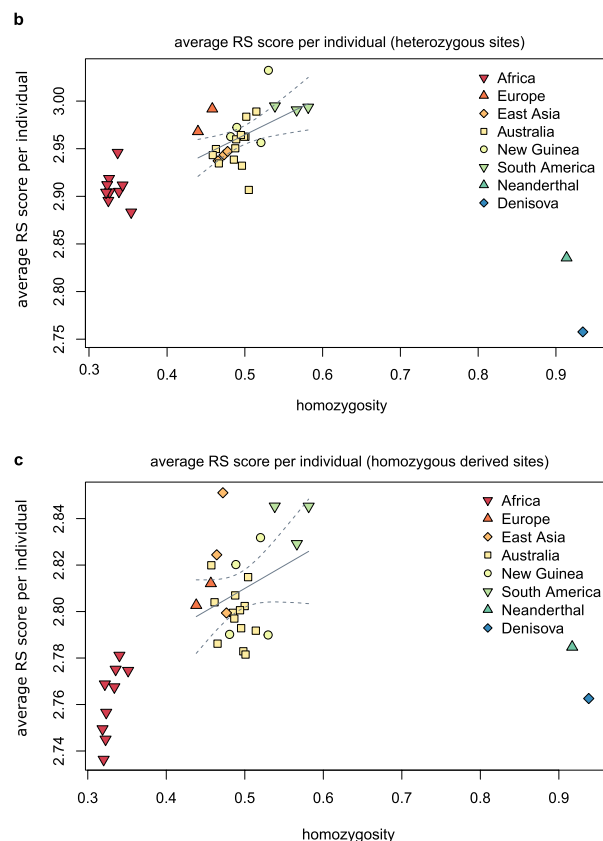
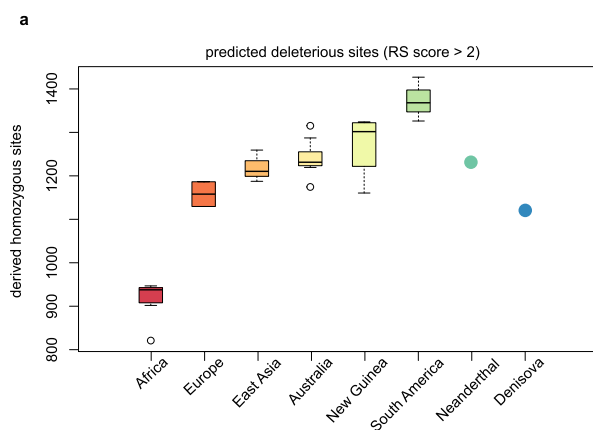
4 kb) of heterozygous or homozygous genotypes in complete linkage disequilibrium ('diplotypes') that are potentially the result of Neanderthal or Denisovan admixture. Those diplotypes are homozygous ancestral in 10 Africans, homozygous derived in the Denisovan for the PDH (respectively Neanderthal for the PNH), homozygous ancestral in the Neanderthal for the PDH (respectively Denisovan for the PNH), and with the derived allele segregating in all other contemporary non-African humans (see Supplementary Information section S10 for details). We report the average number of PDHs and PNHs (**b**), the correlation between the estimated amount of Australo-Papuan ancestry (see Fig. 2a, Extended Data Fig. 1, Supplementary Information section S05) and the number of identified PDHs for each Australian sample (**c**), the sum of the lengths (**d**) and the average length (**e**) of the PDHs and PNHs per individual for worldwide populations included in our reference panel (see Supplementary Information section S04).





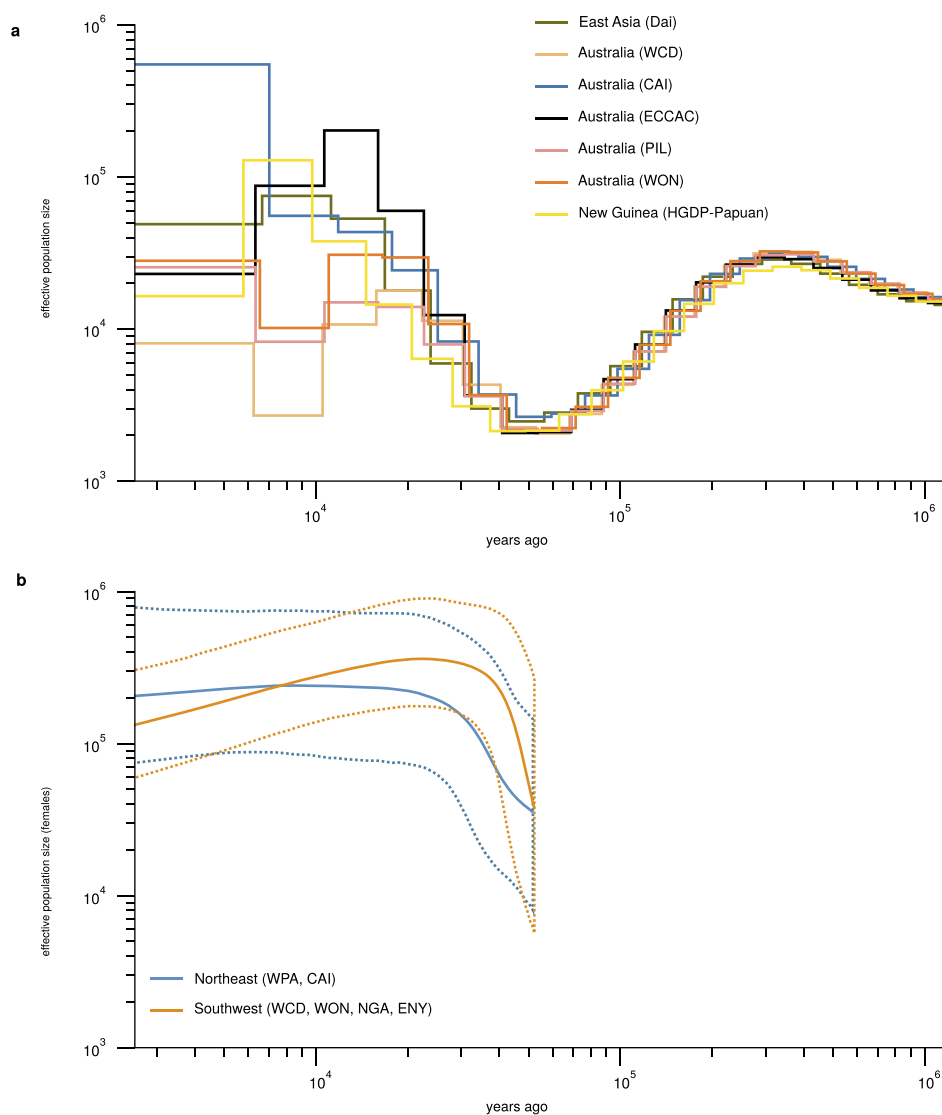
**Extended Data Figure 4 | Out of Africa: admixture graphs based on D-statistics and MSMC analyses.** **a**, Admixture graphs representing some of the topologies considered for the two waves and one wave Out of Africa models assuming Denisovan admixture. All topologies are identical except for the coloured lineages representing Australo-Papuans (green), Neanderthal (Nea, orange) and Denisovan (Den, blue). The graphs differ in (1) the number of OoA events, and (2) the number of Neanderthal admixture pulses. Png, HGDP-Papuan. **b**, Sum of squared errors between the observed D-statistics and the expectations for each quartet in the graph involving the chimpanzee as an outgroup for each of the admixture

graphs shown in **a** and the corresponding four without Denisovan admixture. Each point is the result of the optimization procedure with a different starting point. See Supplementary Information section S09 for details. **c**, Relative cross coalescence rate (CCR) estimates from MSMC<sup>25</sup> for pairs of individuals including one African sample (Yoruba, Dinka and San) and one other, as indicated in the legend. **d**, Simulation study to assess the effect of archaic admixture on the CCR rates. Relative CCR estimated for data simulated under a simple two-population divergence model where one of the populations admixed at different rates with an archaic population. See Supplementary Information section S08 for details.



**Extended Data Figure 5 | Inferred deleterious mutations.** **a**, Box plot of the number of derived homozygous sites per individual for worldwide populations that are predicted to be deleterious. Deleteriousness of SNPs was inferred using genomic evolutionary rate profiling (GERP) rejected substitution scores. Derived alleles with a rejected substitution score larger than 2 were considered to be deleterious, see Supplementary Information section S11. **b**, **c**, Average rejected substitution score per individual calculated across heterozygous sites (**b**), and derived homozygous sites (**c**).

Each coloured symbol corresponds to estimates from a single individual. Homozygosity is calculated as the number of derived homozygous sites divided by the number of sites at which an individual carries at least one copy of the derived allele. Solid lines show the linear regression of homozygosity against average rejected substitution score per individual for non-African modern humans. Dashed lines indicate the 95% confidence interval for the linear regression. See Supplementary Information S11 for details.

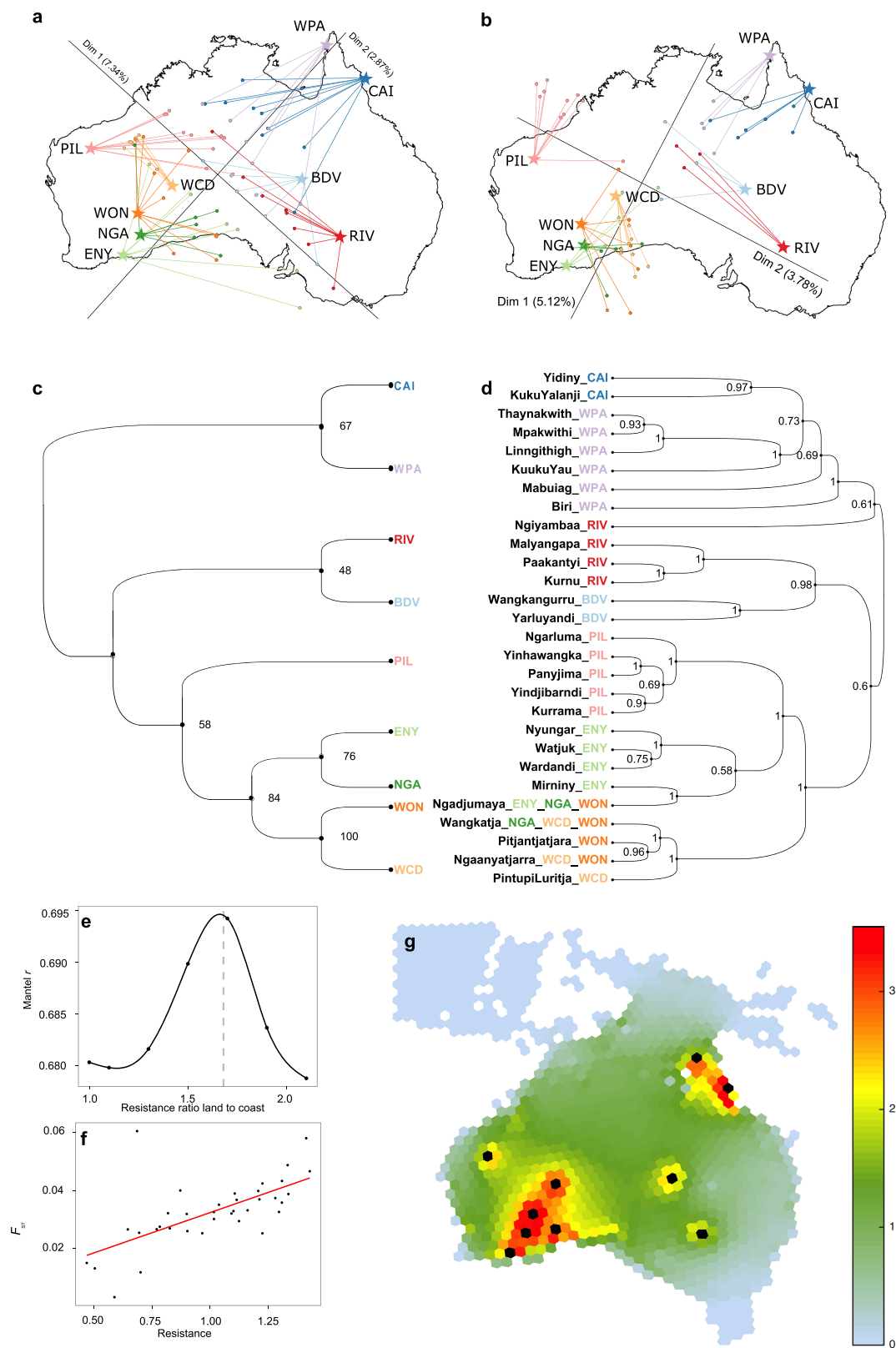


### Extended Data Figure 6 | Effective population size changes over time.

**a**, Population size estimates from MSMC for pairs of individuals from several populations within and outside of Australia. For each run, we used two individuals from each population, that is, four haplotypes in each run. MSMC results were scaled as in Fig. 3. **b**, Bayesian skyline plots (BSP) calculated from the mtDNA genome sequences, showing the effective

population size estimates over time when considering either groups from northeastern Australia (CAI, WPA) or groups from southwestern Australia (ENY, NGA, WCD, WON). Solid lines are the estimates, dashed lines are the corresponding 95% credible intervals (see Supplementary Information section S12).



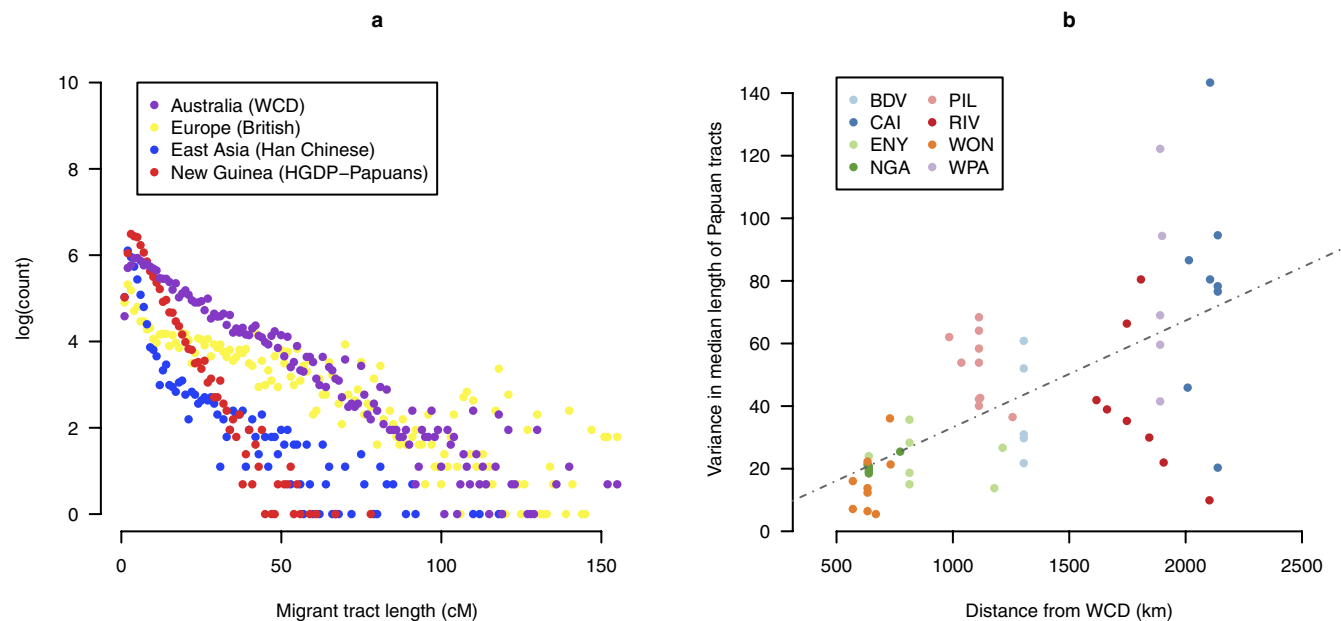


Extended Data Figure 7 | See next page for caption.

**Extended Data Figure 7 | Genetics mirrors geography and languages.**

**a, b**, Procrustes analyses of the first two dimensions of a classical multidimensional scaling (MDS) analysis of the Aboriginal Australian genome sequences (autosomes). We considered two cases: an analysis including all variants (**a**), or only the variants remaining after genomic regions of putative recent European and East Asian origin are ‘masked’ (**b**, Supplementary Information section S06). Both MDS plots have been rotated towards the best overlap with geographic sampling locations as defined by Procrustes analysis<sup>51</sup>. In each plot, the connecting lines indicate the error of the MDS coordinates towards the assigned population-sampling geographic coordinates. We find that the genetic relationships within Australia mirrors geography, with a significant correlation for both cases, that is,  $r_{\text{GEN,GEO}} = 0.59$ ,  $P < 0.0005$  for all variants and even higher,  $r_{\text{GEN,GEO}} = 0.77$ ,  $P < 0.0005$ , for the masked data. We find using the bearing correlogram approach that the main axis of genetic differentiation in the masked Aboriginal Australian genomes is at an angle of  $65^\circ$  compared to the equator, that is, in the southwest to northeast direction (Supplementary Information section S13). **c, d**, Correspondence between genetics and linguistics. **c**, Unrooted neighbour-joining  $F_{\text{ST}}$ -based genetic tree (cladogram). Weir and Cockerham  $F_{\text{ST}}$  distance was computed

between the Aboriginal Australian populations after masking the Eurasian tracts. Statistical robustness of each branch was estimated by means of a bootstrap analysis (1,000 replicates, Supplementary Information section S05). **d**, Bayesian phylogenetic tree for the 28 different Pama–Nyungan languages represented in this sample (from ref. 13, see Supplementary Information section S15). Posterior probabilities are also indicated. Note that one language group can be shared by different Aboriginal Australian groups. The linguistic tree was built with BEAST<sup>52</sup>. **e–g**, Gene flow across the continent. **e**, Mantel non-parametric  $r$  (estimating the goodness of fit between genetic differentiation and connectivity) versus ratios of resistance of inland to coastal nodes, showing a peak at 1.7. **f**, Best fit of pairwise population genetic differentiation,  $F_{\text{ST}}$  (computed between the nine Aboriginal Australian groups after masking Eurasian tracts (Supplementary Information section S06)), versus pairwise connectivity based on the environment (estimated as resistance) when moving inland is 1.7 times harder than moving along coastal nodes. **g**, Gene flow across the Australian landscape, quantified as the cumulative current for pairwise connections among Aboriginal Australian groups (black circles), with larger current (warmer colours) representing greater gene flow.



**Extended Data Figure 8 | European, East Asian and Papuan genomic tracts in Aboriginal Australians.** **a**, Distribution of the tracts assigned to Aboriginal Australian (WCD), Papuan, East Asian or European ancestry for 58 unrelated non-WCD Aboriginal Australian samples. Most of the shorter tracts were of Papuan origin, suggesting that a large fraction of the Papuan gene flow is much older than that from Europe and East Asia, consistent with a Papuan influence spreading slowly from northeastern to southwestern Australia by ancient migration. **b**, Corresponding scatter

plot with fitted line of per-individual variance in Papuan tract length versus geographic distance from WCD, the latter calculated using the great-circle distance formula for pairs of individual GPS coordinates. Papuan tract distribution showed a strong and significant correlation with distance from WCD ( $r=0.64$ ;  $P < 1 \times 10^{-5}$ ), with 'younger tracts' (that is, with a larger variance) closer to New Guinea and 'older tracts' (that is, with a smaller variance) closer to WCD. This is also consistent with continuous Papuan gene flow spreading from the northeast.



**Extended Data Table 1 | Whole genome sequence depth of coverage, haplogroup and language assignments for the Aboriginal Australian samples**

indiv.	DoC*	mtDNA haplotype†	Ychr haplotype‡	Pama-Nyungan language§	indiv.	DoC*	mtDNA haplotype†	Ychr haplotype‡	Pama-Nyungan language§
BDV01	78	S2	-	Yarluyandi Wangkangurru	PIL09	58	S5	R1b1a2a1a2e1	Kurrama
BDV02	75	S1a	R1b1a2a1a2c1g2a1a2	Yarluyandi Wangkangurru	PIL10	61	R	-	Yinhawangka
BDV03	-	-	-	Yarluyandi Wangkangurru	PIL11	57	P3b	C1b	Kurrama
BDV04	70	O1a	-	Yarluyandi Wangkangurru	PIL12	63	P3b	C1b	Yindjibarndi
BDV05	72	S1a	O1a	Yarluyandi Wangkangurru	RIV01	73	M42a	-	Ngiyambaa
BDV06	70	S1a	-	Yarluyandi Wangkangurru	RIV02	62	P4b1	-	Paakantyi
BDV07	70	O1a	-	Yarluyandi Wangkangurru	RIV03	69	M42a	-	Paakantyi
BDV08	70	S1a	R1b1a2a1a2c1g2a1	Yarluyandi Wangkangurru	RIV04	62	P4b1	I2a1a2a1a	Kurnu
BDV09	74	S1a	-	Yarluyandi Wangkangurru	RIV05	72	P4b1	-	Paakantyi
BDV10	72	S1a	I1a2a1d	Yarluyandi Wangkangurru	RIV06	66	H1bs	J2a1b	Ngiyambaa
CAI01	84	P	K2b	Yidiny	RIV07	70	P4b1	R1b1a2a1a2c1c	Paakantyi
CAI02	74	M42	K2b	Yidiny	RIV08	66	P4b1	-	Paakantyi Malyangapa
CAI03	77	M42a	-	Yidiny	WCD01	62	R12	K2b	Ngaanyatjarra
CAI04	71	P	-	Yidiny KukuYalanji	WCD02	59	S1a	C1b	Ngaanyatjarra
CAI05	80	P	O2a1a	Yidiny	WCD03	61	R12	K2b	Wangkatja
CAI06	78	P	C1b	Yidiny	WCD04	52	P3b	K2b	Ngaanyatjarra
CAI07	71	N13	K2b	KukuYalanji	WCD05	60	O1	C1b	Ngaanyatjarra
CAI08	70	P	K2b	Yidiny	WCD06	58	O1a	C1b	Ngaanyatjarra
CAI09	79	P	R1b1a2a1a2b1	Yidiny	WCD07	61	M42	-	Ngaanyatjarra
CAI10	73	E1a2	K2b	-	WCD08	64	M42	-	Ngaanyatjarra
ENY01	69	H1e1a3	R1b1a2a1a2b1c1	Nyungar	WCD09	59	R	J2a1b	Ngaanyatjarra
ENY02	79	R12	-	Ngadjumaya	WCD10	63	M42	-	Ngaanyatjarra
ENY03	83	O	-	Miriny	WCD11	57	M42	K2b	Ngaanyatjarra
ENY04	83	M42	-	Nyungar	WCD12	59	M42	C1b	Ngaanyatjarra PintupiLuritja
ENY05	78	S2	-	Ngadjumaya	WCD13	67	M14	C1b	Ngaanyatjarra
ENY06	70	M42	-	Wardandi	WON01	71	O	I1a2a1a3a	Wangkatja
ENY07	73	S2	E1b1b1b2a	Watjuk	WON02	101	O1a	-	Wangkatja
ENY08	71	P4b1	C1b	Nyungar Ngadjumaya	WON03	65	O1a	-	Wangkatja
NGA01	74	O1	-	Ngadjumaya	WON04	58	R	-	Ngaanyatjarra
NGA02	52	O1a	-	Ngadjumaya	WON05	56	O1a	I2a2a1a2a2	Wangkatja
NGA03	73	O	-	Ngadjumaya	WON06	60	R12	-	Wangkatja
NGA04	75	O	R1b1a2a1a1b1a1a	Wangkatja	WON07	57	O	-	Ngadjumaya
NGA05	56	R12	-	Ngadjumaya	WON08	52	O	-	Wangkatja
NGA06	63	S1a	-	Wangkatja	WON09	20	O	E1b1b1a1b1a4	Wangkatja
PIL01	58	R	C1b	Yinhawangka	WON10	50	O1	R1b1a2a1a2a	Wangkatja
PIL02	61	M42	C1b	Yinhawangka	WON11	58	R12	-	Pitjantjatjara
PIL03	56	M42	-	Yinhawangka	WPA01	51	P5	-	Thaynakwith Linglithigh
PIL04	64	M42	-	Yinhawangka	WPA02	50	P	C1b	Mpakwithi Kaanju
PIL05	68	M42	C1b	Yinhawangka	WPA03	51	M42a	K2b	Thaynakwith Biri
PIL06	59	O1	K2b	Panyjima	WPA04	52	P5	-	Thaynakwith KukuYau
PIL07	63	O	-	Panyjima	WPA05	56	M42	NA	Mabuiag Thaynakwith
PIL08	72	M42	C1b	Yindjibarndi Kurrama	WPA06	53	P5	O1a	Mpakwithi

\*The depth of coverage (DoC) is the average number of reads covering every position in the genome (hg19) after duplicate removal (see Supplementary Information section S03).

†The average depth of coverage on the mitochondrial genome (mtDNA) is 3,484 ± 1,515 (mean ± s.d.) and haplogroups were called with haplogrep (<http://haplogrep.uibk.ac.at/>) and haplofind (<https://haplofind.unibo.it/>), see Supplementary Information section S12 for details and references.

‡The average depth of coverage on the Y chromosome (Ychr) is 28.9 ± 4.5 (mean ± s.d.). Haplogroup assignment was performed with an in-house script that matched our SNPs with the classification provided in ISOGG version 10.08, see Supplementary Information section S12 for details and references.

§Language group with which the speaker self-identifies, or to which they were assigned. Where more than one language is given, speakers either identified with more than one group, or they could not be assigned to a single group with certainty.

Extended Data Table 2 | Selection scan in Aboriginal Australians

Focal Pop	Nearby Gene*	Position†	rsID	Dist ‡	PBSn1§	F <sub>12</sub> ¶	F <sub>13</sub>	F <sub>23</sub>	Function of gene product#
All	<i>TMEM86B</i>	55,833,076	rs734517	92,444	0.78	0.93	0.99	0.06	Catalyzes the degradation of lysoplasmalogen. Modulates cell membrane proteins.
All	<i>LRR52</i>	165,621,695	rs4147601	88,510	0.74	0.96	0.91	0.01	Modulates voltage of potassium ion channels. Expressed in testis.
All	<i>MACROD2</i>	15,209,684	rs175279	901	0.70	0.92	0.89	-0.01	Involved in deacetylase activity. Possibly (but not conclusively) causative of Kabuki syndrome.
All	<i>JRKL</i>	96,747,146	rs72959058	507,105	0.74	0.99	0.87	0.15	Homologue to "jerky" gene in mouse.
All	<i>SPATA20</i>	48,631,324	rs73338243	287	0.70	0.96	0.85	0.09	Spermatid protein.
All	<i>NAA60</i>	3,537,933	rs73503305	970	0.71	0.91	0.91	-0.02	Histone acetyltransferase required for nucleosome assembly and chromosome segregation during anaphase. Human-specific imprinted gene.
All	<i>CBLN2</i>	70,019,066	rs12455116	184,848	0.69	0.92	0.87	0.00	<i>CBLN2</i> : cerebellum-specific protein involved in various signaling pathways. Possibly associated with pulmonary arterial hypertension.
	<i>NETO1</i>			390,482					<i>NETO1</i> : brain-specific transmembrane protein involved in the regulation of neuronal circuitry. Associated with thyroid function.
All	<i>SLC2A12</i>	134,391,056	rs4896021	17,267	0.76	0.96	0.95	-0.01	Catalyzes sugar absorption. Involved in the pathogenesis of diabetes. Associated with serum urate levels.
All	<i>LOC101927657</i>	127,358,509	rs145200081	16,731	0.65	0.94	0.80	0.13	Unknown (ncRNA).
All	<i>LOC102724612</i>	64,466,486	rs113341339	78,446	0.73	0.91	0.95	0.00	Unknown (ncRNA).
NE	<i>ZBTB20</i>	114,530,679	rs9289004	10,658	0.55	0.65	0.82	0.07	Transcriptional repressor associated with Primrose syndrome.
NE	<i>ANXA10</i>	168,646,016	rs2176513	367,671	0.49	0.61	0.61	-0.01	Calcium-dependent phospholipid-binding annexin.
NE	<i>TRPC3</i>	122,905,041	rs4502701	32,132	0.50	0.59	0.64	-0.01	Non-selective cation channel, associated with spinocerebellar ataxia.
NE	<i>HS3ST1</i>	11,634,592	rs7665516	204,055	0.45	0.45	0.71	0.07	Regulates rate of generation of anticoagulant heparan sulfate proteoglycan.
NE	<i>MIR548C</i>	65,027,511	rs2620721	11,126	0.50	0.55	0.73	0.03	Unknown (microRNA).
NE	<i>STARD13</i>	33,799,901	rs7318080	19,714	0.49	0.54	0.83	0.20	Involved in cell proliferation and fibroblast morphology.
NE	<i>AKAP11</i>	42,931,386	rs7319267	33,983	0.53	0.56	0.85	0.13	Directs protein kinase A activity and is involved in cAMP messenger signaling.
NE	<i>AGMO</i>	15,212,231	rs35557899	27,711	0.47	0.51	0.68	0.01	Catalyzes the cleavage of O-alkyl bonds of ether lipids.
NE	<i>RUNX1T1</i>	92,925,296	rs11776341	41,898	0.45	0.56	0.54	0.00	Involved in transcriptional repression. A translocation involving this gene is associated with acute myeloid leukemia.
NE	<i>FHAD1</i>	15,680,451	rs2473358	971	0.45	0.60	0.52	0.00	Unknown.
SW	<i>KCNJ2</i>	68,190,552	rs35167900	14,369	0.57	0.61	0.93	0.22	Potassium channel, associated with familial atrial fibrillation and periodic paralysis.
SW	<i>TACC2</i>	123,754,065	rs10159998	5,062	0.50	0.60	0.67	0.00	Belongs to a family of proteins that interact with the centrosome and microtubules, and that are implicated in cancer.
SW	<i>LOC101928708</i>	87,228,164	rs4843556	17,556	0.58	0.65	0.86	0.07	Unknown (ncRNA).
SW	<i>C16orf82</i>	27,187,689	rs72782349	107,202	0.51	0.60	0.69	0.02	Unknown.
SW	<i>LOC100507391</i>	194,520,805	rs56379930	17,908	0.55	0.66	0.75	-0.01	Unknown (ncRNA).
SW	<i>HAUS4</i>	23,416,252	rs2008951	127	0.49	0.50	0.83	0.16	A component of a microtubule-binding complex that plays a role in the generation of microtubules in the mitotic spindle.
SW	<i>KNG1</i>	186,438,819	rs5029990	815	0.51	0.56	0.72	0.01	During the inflammatory response, it is involved in vasodilation, coagulation, enhanced capillary permeability and pain induction.
SW	<i>MYDGF</i>	4,657,016	rs66891175	540	0.55	0.61	0.88	0.16	Unknown.
SW	<i>MSMP</i>	35,757,075	rs1951432	2,801	0.48	0.47	0.88	0.27	May be involved in the tumorigenesis of prostate cancer.
SW	<i>VAV2</i>	136,756,316	rs2519771	29,762	0.47	0.51	0.73	0.07	Member of an oncogene family. Involved in T-cell receptor signaling.

Top 10 peaks of differentiation from genome scans of all Aboriginal Australians combined (All) and two Aboriginal Australians subgroups living in different ecological regions in Australia, the northeast (NE) or southwest (SW).

\*RefSeq protein coding gene with exon boundary near to windowed-PBSn1 peak.

†Genomic position (hg19) of SNP with highest value of PBSn1 within 200 Mb of the top window.

‡Distance between SNP and the nearest exon boundary of nearest gene.

§PBSn1 statistic at top SNP.

¶F<sub>ST</sub> statistics at top SNP for each comparison within the PBSn1 calculation.

#Please see Supplementary Information section S16 for references.

# Quantifying crater production and regolith overturn on the Moon with temporal imaging

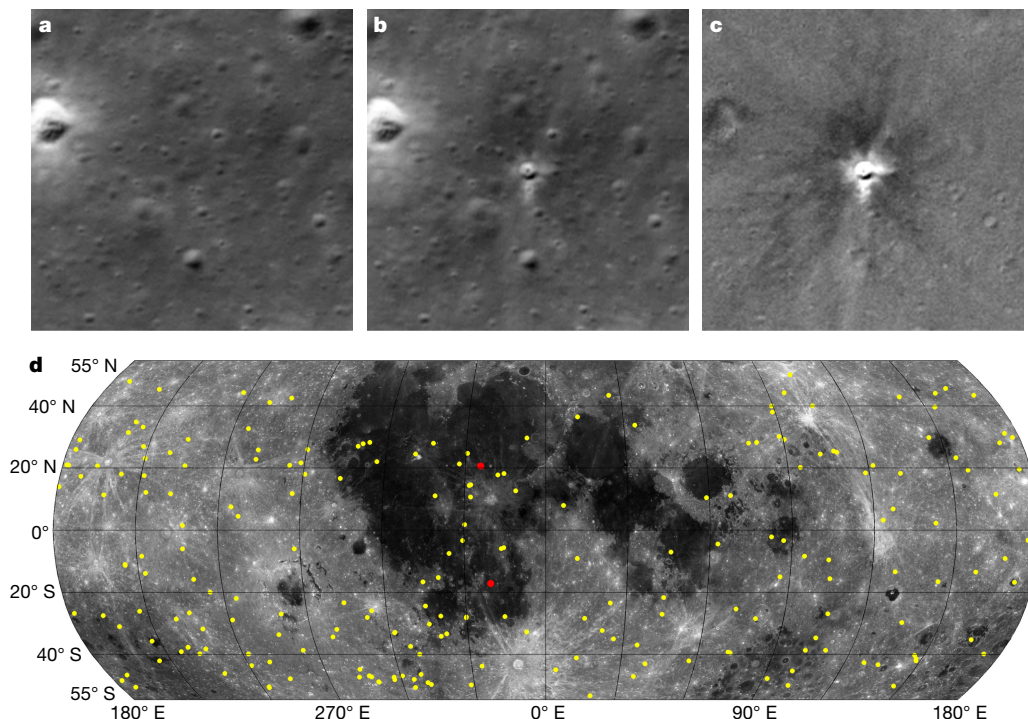
Emerson J. Speyerer<sup>1</sup>, Reinhold Z. Povilaitis<sup>1</sup>, Mark S. Robinson<sup>1</sup>, Peter C. Thomas<sup>2</sup> & Robert V. Wagner<sup>1</sup>

Random bombardment by comets, asteroids and associated fragments form and alter the lunar regolith and other rocky surfaces. The accumulation of impact craters over time is of fundamental use in evaluating the relative ages of geologic units. Crater counts and radiometric ages from returned samples provide constraints with which to derive absolute model ages for unsampled units on the Moon and other Solar System objects<sup>1–4</sup>. However, although studies of existing craters and returned samples offer insight into the process of crater formation and the past cratering rate, questions still remain about the present rate of crater production, the effect of early-stage jetting during impacts and the influence that distal ejecta have on the regolith. Here we use Lunar Reconnaissance Orbiter Camera (LROC) Narrow Angle Camera (NAC) temporal ('before and after') image pairs to quantify the contemporary rate of crater production on the Moon, to reveal previously unknown details of impact-induced jetting, and to identify a secondary impact process that is rapidly churning the regolith. From this temporal dataset, we detected 222 new impact craters and found 33 per cent more craters (with diameters of at least ten metres) than predicted by the standard

Neukum production and chronology functions for the Moon<sup>2</sup>. We identified broad reflectance zones associated with the new craters that we interpret as evidence of a surface-bound jetting process. We also observe a secondary cratering process that we estimate churns the top two centimetres of regolith on a timescale of 81,000 years—more than a hundred times faster than previous models estimated from meteoritic impacts (ten million years)<sup>5</sup>.

Temporal imaging provides a means to directly identify, quantify and characterize new impact craters and other contemporary surface changes. Using metre-scale NAC observations, five new craters were identified<sup>6</sup> by comparing NAC images to Apollo Panoramic images. Recent impact flashes observed from Earth aided in the discovery of new craters with diameters of 18 m and 34 m (refs 7–9). Building on these findings and the growing imaging time base, we investigated 14,092 NAC temporal pairs (see Methods) covering  $2.49 \times 10^6$  km<sup>2</sup> (6.6% of the lunar surface) and detected changes that occurred between the observations (Fig. 1a–c).

From this analysis, we discovered 222 newly formed impact craters with diameters between 43 m and down to the resolution limit of the

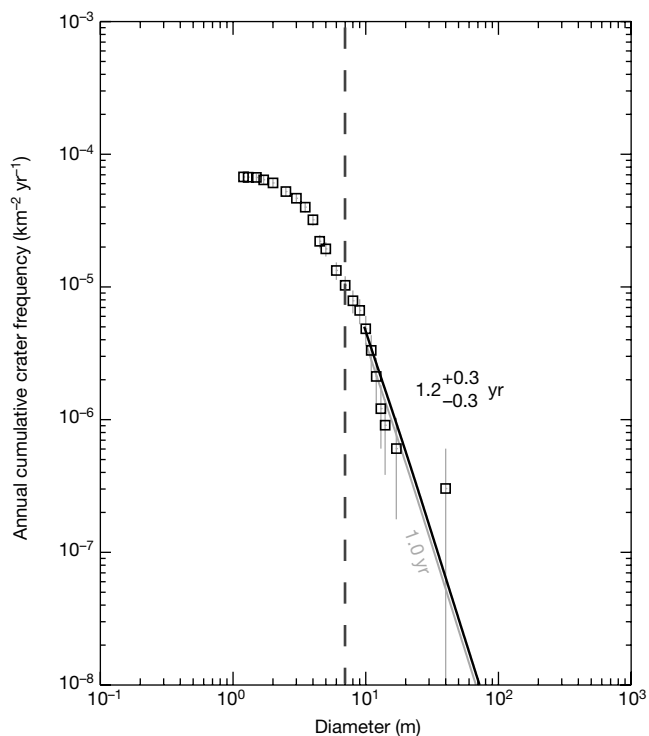


**Figure 1 | Detection and distribution of new impact craters.** **a, b**, Cropped example of a temporal pair (**a**, before image; **b**, after image) showing a 14.3-m impact crater (NAC frames M1104273380R and M1180855200R). **c**, Temporal ratio image created by dividing the after

image (**b**) by the before image (**a**). **d**, Distribution of 222 new craters discovered with temporal pairs (yellow markers). The red markers identify the two craters located with impact flashes. The region shown in **a–c** is 225 m across.

<sup>1</sup>Arizona State University, School of Earth and Space Exploration, Tempe, Arizona 85287, USA. <sup>2</sup>Cornell University, Cornell Center for Astrophysics and Planetary Science, Ithaca, New York 14853, USA.





**Figure 2 | Annual cumulative size–frequency distribution of newly formed craters discovered with NAC temporal pairs.** Owing to the identification of more craters than the NPF predicts, the NPF model age of the surface covered by temporal pairs is estimated to be  $1.3 \pm 0.3$  yr (fit not shown), or  $1.2 \pm 0.3$  yr when excluding the largest crater (43 m) from the fit. For reference, a 1-yr isochron (grey line) and a fit using a Poisson calculation<sup>10</sup> (black line) derived from the NPF is overlaid. The dashed vertical line represents the completeness limit for the current collection of temporal pairs. Error bars correspond to Poisson statistics.

image pairs (about four pixels) (Fig. 1d). Using the statistics of the new crater sizes and a normalized surface area, we derived an initial, annual size–frequency distribution of the craters to estimate the non-saturated, contemporary crater-production rate (Fig. 2). This rate is compared to estimates derived from the Neukum production function (NPF) and corresponding chronology function, which is a commonly used model for estimating the formation rate of craters as small as 10 m and is used to assign model age estimates to geologic units<sup>2</sup>. Although pixel scale limits the measured crater distribution at small diameters ( $<7$  m), we find 33% more craters with diameters of  $\geq 10$  m than the NPF predicts ( $16 \text{ yr}^{-1}$  in NAC temporal pairs versus  $12 \text{ yr}^{-1}$  predicted by the NPF over the same surface area) and a model age<sup>10</sup> derived from the NPF<sup>2</sup> that is 20% older than predicted. Given the modelled production rate of new craters<sup>2</sup> and our search area, there is only a 16% chance of identifying 16 craters, which implies a possibly higher production rate for craters with diameters of  $\geq 10$  m. Furthermore, in the 10–20-m diameter range, the power-law slope ( $-b$ ) of the cumulative crater frequency of our initial crater dataset is steeper than the NPF predicts<sup>2</sup> ( $b_{\text{NAC}} = 4.64$ , 95% confidence interval (CI) of 4.15–5.14 versus  $b_{\text{NPF}} = 2.97$ –3.15). The steeper slope was also present when analysing the differential crater frequency over the same diameter range. The tentatively steeper slope, which is also seen in extrapolated cratering models derived from the known orbital and size distribution of asteroids and comets<sup>11</sup>, implies a higher crater-production rate for smaller impactors than the NPF estimates (see Methods).

Mapping of the individual impact sites reveals evidence of the formation process in the form of up to four distinct reflectance zones: proximal high reflectance zone (PHRZ), proximal low reflectance zone (PLRZ), distal high reflectance zone (DHRZ) and distal low reflectance zone (DLRZ). These reflectance zones are the result of ballistic and

jetted material modifying surface properties. The four reflectance zones were first documented<sup>7</sup> around an 18-m impact crater formed on 17 March 2013 (event recorded by Earth-based assets)<sup>8</sup>.

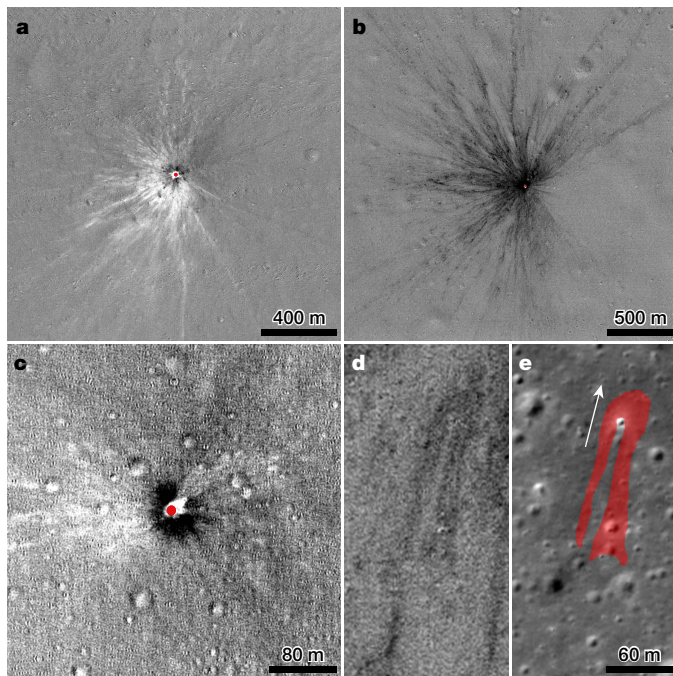
From our analysis, we found that the proximal zones fall within several crater diameters  $D$  from the crater centre. In the 58 cases for which a PHRZ was identified, it was typically the closest reflectance zone to the impact crater and modified the surface reflectance (relative to the pre-impact surface) by  $12\% \pm 9\%$ , and extended from the crater centre an average of  $2.1D$  (median of  $1.5D$ ). The PLRZ, which was present in nearly all cases, exhibited a lower surface reflectance ( $-9\% \pm 4\%$  on average) when compared to pre-existing terrain, and extended an average of  $3.2D$  from the centre of the crater regardless of the presence of a PHRZ (median of  $2.9D$ ).

Beyond the proximal zone(s), we always found at least one distal reflectance zone. The DLRZ ( $-2\%$  to  $-6\%$  reflectance change) is visible around 84% of the new impacts. In addition, 24% of the new craters (including the 17 March 2013 crater<sup>7</sup>) have a DHRZ (1% to 4% higher reflectance). The DLRZ extends out from the proximal zone(s) in a ray-like pattern to a distance that is proportional to the area of the source crater ( $5.03D^{2.0}$ ), whereas the DHRZ extends  $0.52D^{2.5}$  from the centre of the crater, with  $D$  and the solution in units of metres (see Methods). In some cases, such as the 17 March 2013 impact, the DHRZ is consistently closer to the crater than the DLRZ<sup>7</sup>; but we found that the order and placement varies from crater to crater (Fig. 3a, b).

Both proximal zones around the 17 March crater are considered<sup>7</sup> to contain the continuous ejecta blanket from the impact, with the albedo change depending on the relative optical maturity of the ejecta<sup>12</sup> or changes to the local surface roughness. Our observations are consistent both with that interpretation and with the observation that only a subset of the impacts reached an optically immature layer to produce a PHRZ (Fig. 3c). The origins of the distal zones are not as clear. The DHRZ has been attributed<sup>7</sup> to impact-induced jetting, and the DLRZ to sparse ballistic sedimentation from the later ejecta curtain churning the local regolith<sup>7,13,14</sup>. Further analysis of the new craters reported here suggests that jetting may have played a dominant part in forming both distal zones.

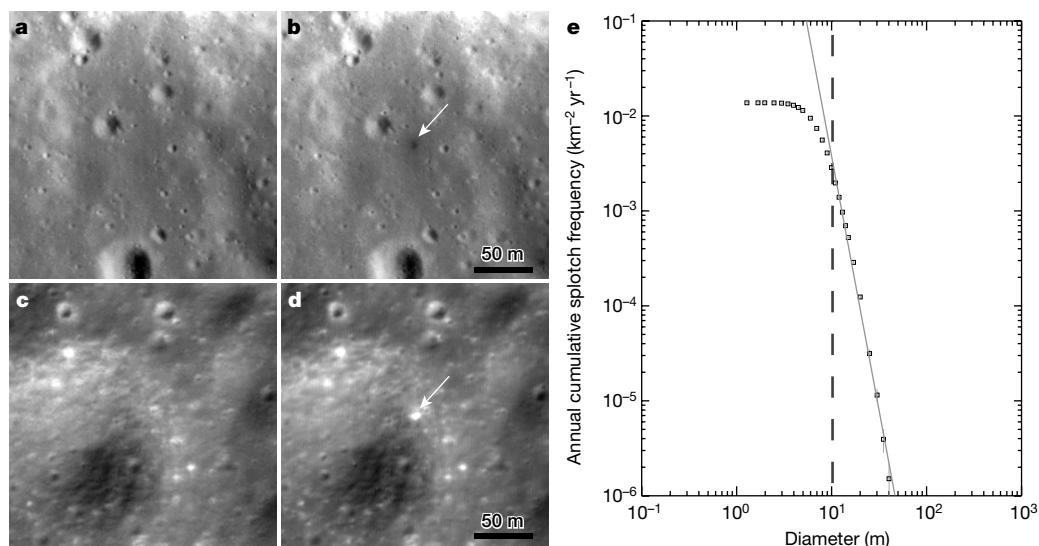
Jetting occurs early in the crater-formation process when the shock wave produces a mixture of melted and vaporized material that is ejected at low angles and at extremely high speeds that can exceed the original speed of the impactor<sup>13,15</sup>. We agree with the assertions of ref. 7 that jetted vapour may smooth and redistribute surface grains, a process that could destroy the highly porous (up to 90%) structure in the upper few centimetres of regolith<sup>16</sup> and form the DHRZ. In addition, we propose that jetting could locally increase the surface roughness, resulting in the DLRZ. Small topographic features (for example, crater rims) have ‘shadows’ in the distal zones (DHRZ and DLRZ), which indicates that these zones formed from an effect travelling parallel to and along the surface (Fig. 3d, e), rather than from a classic ballistic-sedimentation process. The pressure gradient between the jet and the vacuum of space causes the jet to expand and interact with the surface for tens to hundreds of crater diameters. To create the DLRZ, we propose that jetted melt and/or fine-grained regolith carried by the jetted vapour creates a series of impacts downrange that churn the upper several millimetres of regolith, which increases the surface roughness and creates macroscopic shadowing that reduces the reflectance<sup>7,17,18</sup>. This localized churning caused by the jetting may also increase the pore spacing in the upper regolith and form the thermo-physical cold spots observed around young impacts<sup>19</sup>.

We attribute the variation in the production of two distal zones to the proportions of vapour and melt (or other clastic material) within the jet itself<sup>20,21</sup>. It has been proposed<sup>20</sup> that the initial and most highly shocked portion of the jet contains more vapour, whereas later-stage jetting tends to consist of more melt than vapour. Although the vapour portion of the jet may be able to effectively smooth the porous regolith similar to a blast zone<sup>22–24</sup>, the increased mass of the melted material in the melt-dominated portion of the jet can collapse to the surface and



**Figure 3 | Signatures of crater formation in NAC temporal-pair ratios.** **a**, Temporal ratio image showing rays of a DHRZ from a 19.5-m crater (ratio of NAC frames M1132496422R/M1101866147R). **b**, Temporal ratio image displaying a DLRZ extending 150D from a 12-m impact crater (ratio of NAC frames M1105837846R/M1121160416R). **c**, Temporal ratio image showing the distinct boundary between the two proximal and distal zones around an 11-m crater (ratio of NAC frames M1113630818R/M1098309025L). The red dots in **a–c** mark the location of the crater. **d**, Enlarged version of **b** showing 'shadows' in the DLRZ caused by a small crater. **e**, NAC frame M1132496422R highlighting the DLRZ shadow shown in **d** with an arrow pointing to parent impact.

churn the upper few millimetres of regolith. A lack of a DHRZ around some craters may indicate that the jetted material was less powerful or carried more fines and/or melt. Furthermore, we see evidence that the impact angle and local topography affects the resulting albedo patterns (Fig. 3).



**Figure 4 | Examples and annual size–frequency distribution of splotches.** **a**, **b**, Temporal pair (**a**, before image; **b**, after image) showing a low-reflectance splotch (indicated by the arrow in **b**) (NAC frames M1190544405R and M1182416597L). **c**, **d**, Temporal pair (**c**, before image; **d**, after image) showing a high-reflectance splotch (indicated by the arrow

In addition to surface modification adjacent to the new craters, NAC temporal pairs revealed over 47,000 other changes in surface reflectance. These changes in reflectance lacked rims, or any other resolvable topographic signature. We interpret these localized reflectance changes to be the result of small primary or secondary impacts and refer to them as “splotches”<sup>7</sup> (Fig. 4a–d). Most (90.7%) of the splotches exhibit lower reflectance (average  $-4\%$ ) than did the same region before the splotches formed (Extended Data Fig. 1), whereas 7.4% show an increase in surface reflectance (average  $+10\%$ ) and the remaining 1.9% have mixed reflectance patterns. Many of the splotches are nearly circular with diameters ranging from 30 m down to the detection limit (about five pixels) (Fig. 4e). On the basis of the number of observed splotches, we estimate that  $1.09 \times 10^5$  splotches with diameters of  $>10$  m form annually on the Moon and that they alone can alter 99% the surface over  $1.0 \times 10^7$  yr. Using a power-law fit (see Methods) and extrapolating the results to 1 m yields an annual rate of  $40.2 \pm 1.6$  splotches per square kilometre with a 99% coverage rate after  $8.1 \times 10^4$  yr.

Although some of the splotches may contain unresolved primary impact craters ( $D < 4$  pixels), we propose that a majority are the result of secondary impacts of poorly consolidated (or loose) regolith that only churned the target surface and do not form a classic impact crater. Localized groups of splotches observed around new craters are consistent with this interpretation. For example, the 17 March 2013 impact site contained 248 splotches within 30.3 km of the parent crater<sup>7</sup>. In addition, some splotches exhibit an asymmetric arrow-head shape or herringbone pattern that points back to new impact craters<sup>7</sup>.

Such secondary impacts of loose clusters of regolith create a hummocky and pitted surface surrounded by a subdued rim with a ratio of depth to diameter of 1:30 (ref. 25). This inner morphology is similar to the widespread “raindrop” texture described by the Apollo astronauts<sup>26</sup>. Given such shallow excavation, most of the splotches would churn or garden only the upper few centimetres of regolith (mature zone), which would have similar optical maturity properties to the regolith at the surface. This shallow excavation is consistent with our observations, which indicate that high-reflectance splotches are typically present in areas where a thin layer of mature regolith might be expected (that is, on slopes of  $>15^\circ$  and ejecta of Copernican-aged craters; Fig. 4c, d). Applying a conservative estimate of churning depth of 1:50 to the identified splotches, we predict that the top 20 cm of regolith will

in **d**) on a crater wall (NAC frames M1104387952R and M1180961488L). **e**, Annual cumulative size–frequency distribution of new splotches derived with temporal pairs. The grey line is a power-law fit used to estimate the diameters of splotches finer than the detection resolution at all pixel scales ( $<10$  m). Error bars correspond to Poisson statistics.



be altered over  $1.0 \times 10^7$  yr by  $\geq 10$ -m splotches. Extrapolating the splotch population down to  $\geq 1$  m, our model predicts that only about  $8.1 \times 10^4$  yr is required to rework and garden the upper 2 cm of regolith, which corresponds to a rate that is more than 100 times faster than previous models predicted from meteoritic impacts ( $10^7$  yr)<sup>5</sup>. This rapid churning is consistent with Apollo drive core samples, in which the top 10–50 cm of regolith is locally reworked and homogenized<sup>27</sup>, and measurements of short-lived cosmogenic radionuclides (<sup>26</sup>Al, half-life  $t_{1/2} = 7.3 \times 10^4$  yr) that indicate that the upper 2–3 cm has been continuously reworked over a period of  $10^5$ – $10^6$  yr (refs 28–30).

This rapid, localized, vertical churning affects the final thickness of the mature top layer of regolith, spurs volatile redistribution, accelerates the degradation of surface features and poses a minor hazard to future long-lived human and robotic surface assets. The impressive population of splotches compared to the number of new craters (47,000 versus about 220) stresses the importance of secondary impact processes on the lunar surface. We found that the 1-yr isochron derived from the NPF<sup>2</sup> is within the estimated uncertainty of our contemporary crater-production rate, which is consistent with the notion that the cratering rate has been uniform over recent geologic time. However, the discovery, in the limits of the available data, of more craters (33% more with  $D \geq 10$  m) and a potentially steeper cumulative slope in the 10–20-m diameter range suggests that the cratering rate in this size range has not been fully characterized in previous models and might not account for crater degradation caused by splotches and other impacts. Further temporal observations will reduce the uncertainty estimates and better describe the current accumulation rate of craters with diameters of 100 m and smaller.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 8 February; accepted 1 September 2016.**

- Ivanov, B. A. in *Chronology and Evolution of Mars* (eds Kallenbach, R. et al.) 87–104 (Springer, 2001).
- Neukum, G., Ivanov, B. A. & Hartmann, W. K. Cratering records in the inner Solar System in relation to the lunar reference system. *Space Sci. Rev.* **96**, 55–86 (2001).
- Hartmann, W. K. Martian cratering 8: isochron refinement and the chronology of Mars. *Icarus* **174**, 294–320 (2005).
- Stöffler, D. & Ryder, G. Stratigraphy and isotope ages of lunar geologic units: chronological standard for the inner Solar System. *Space Sci. Rev.* **96**, 9–54 (2001).
- Gault, D. E., Hörz, F., Brownlee, D. E. & Hartung, J. B. Mixing of the lunar regolith. In *Proc. Lunar Sci. Conf. 5th Vol.* 3 (ed. Gose, W. A.) 2365–2386 (Lunar and Planetary Institute, 1974).
- Daubar, I. J. et al. New craters on Mars and the Moon. *Lunar Planet. Sci. Conf.* **42**, abstr. 2232 (2011).
- Robinson, M. S. et al. New crater on the Moon and a swarm of secondaries. *Icarus* **252**, 229–235 (2015).
- Suggs, R. M., Moser, D. E., Cooke, W. J. & Suggs, R. J. The flux of kilogram-sized meteoroids from lunar impact monitoring. *Icarus* **238**, 23–36 (2014).
- Madiedo, J. M., Ortiz, J. L., Morales, N. & Cabrera-Cano, J. A large lunar impact blast on 2013 September 11. *Mon. Not. R. Astron. Soc.* **439**, 2364–2369 (2014).
- Michael, G. G., Kneissl, T. & Neesemann, A. Planetary surface dating from crater size–frequency distribution measurements: Poisson timing analysis. *Icarus* **277**, 279–285 (2016).
- Le Feuvre, M. & Wieczorek, M. A. Nonuniform cratering of the Moon and a revised crater chronology of the inner Solar System. *Icarus* **214**, 1–20 (2011).
- Shoemaker, E. M. in *Physics and Astronomy of the Moon* (ed. Kopal, Z.) 283–357 (Academic Press, 1963).
- Melosh, H. J. *Impact Cratering: A Geologic Process* (Oxford Univ. Press, 1989).
- Oberbeck, V. R. The role of ballistic erosion and sedimentation in lunar stratigraphy. *Rev. Geophys. Space Phys.* **13**, 337–362 (1975).
- Johnson, B. C., Bowling, T. J. & Melosh, H. J. Jetting during vertical impacts of spherical projectiles. *Icarus* **238**, 13–22 (2014).
- Hapke, B. & van Hoen, H. Photometric studies of complex surfaces, with applications to the Moon. *J. Geophys. Res.* **68**, 4545–4570 (1963).
- Hapke, B. Bidirectional reflectance spectroscopy. *Icarus* **195**, 918–926 (2008).
- Veverka, J. in *Physical Studies of Minor Planets* NASA SP-267 (ed. Gehrels, T.) 79–90 (NASA, 1971).
- Bandfield, J. L. et al. Lunar cold spots: granular flow features and extensive insulating materials surrounding young craters. *Icarus* **231**, 221–231 (2014).
- Vickery, A. M. The theory of jetting: application to the origin of tektites. *Icarus* **105**, 441–453 (1993).
- Melosh, H. J. & Sonett, C. P. When worlds collide: jetted vapor plumes and the Moon's origin. In *Origins of the Moon Proc. Conf.* (eds Hartmann, W. K. et al.) 621–642 (Lunar and Planetary Institute, 1986).
- Clegg, R. N., Jolliff, B. L., Robinson, M. S., Hapke, B. W. & Plescia, J. B. Effects of rocket exhaust on lunar soil reflectance properties. *Icarus* **227**, 176–194 (2014).
- Clegg-Watkins, R. N. et al. Photometric characterization of the Chang'e-3 landing site using LROC NAC images. *Icarus* **273**, 84–95 (2016).
- Shkuratov, Y., Kaydash, V., Sysolyatina, X., Razim, A. & Videen, G. Lunar surface traces of engine jets of Soviet sample return probes: the enigma of the Luna-23 and Luna-24 landing sites. *Planet. Space Sci.* **75**, 28–36 (2013).
- Schultz, P. H. & Gault, D. E. Clustered impacts: experiments and implications. *J. Geophys. Res.* **90**, 3701–3732 (1985).
- Swann, G. A. et al. *Geology of the Apollo 14 Landing Site in the Fra Mauro Highlands*. Professional Paper 880 (US Geological Survey, 1977).
- McKay, D. S. et al. in *Lunar Sourcebook* (eds Heiken, G. H. et al.) 285–356 (Cambridge Univ. Press, 1991).
- Fruchter, J. S., Rancitelli, L. A. & Perkins, R. W. Recent and long-term mixing of the lunar regolith based on <sup>22</sup>Na and <sup>26</sup>Al measurements in Apollo 15, 16, and 17 deep drill stems and drive tubes. In *Proc. Lunar Sci. Conf. 7th Vol.* 1 (ed. Merrill, R. B.) 27–39 (Lunar Planetary Institute, 1976).
- Fruchter, J. S., Rancitelli, L. A., Laul, J. C. & Perkins, R. W. Lunar regolith dynamics based on analysis of the cosmogenic radionuclides <sup>22</sup>Na, <sup>26</sup>Al, and <sup>53</sup>Mn. In *Proc. Lunar Sci. Conf. 8th Vol.* 3 (ed. Merrill, R. B.) 3595–3605 (Lunar Planetary Institute, 1977).
- Fruchter, J. S., Rancitelli, L. A., Evans, J. C. & Perkins, R. W. Lunar surface processes and cosmic ray histories over the past several million years. In *Proc. Lunar Sci. Conf. 9th Vol.* 2 (ed. Merrill, R. B.) 2019–2032 (Lunar Planetary Institute, 1978).

**Acknowledgements** We acknowledge the engineers and technical support team at NASA Goddard Space Flight Center and Arizona State University who enable the collection of a vast image archive of the lunar surface that will be used for decades to come. This work is supported by the Lunar Reconnaissance Orbiter (LRO) Project and the Arizona State University LROC contract.

**Author Contributions** E.J.S. drafted the manuscript and authored the CRISP software used to identify surface changes. R.Z.P. classified and catalogued the temporal changes. M.S.R. is the principal investigator for the Lunar Reconnaissance Orbiter Camera and provided key contributions to the scientific interpretations. P.C.T. aided in the scientific interpretations of splotches and reflectance zones. R.V.W. assisted in optimizing the change detection software and assessed temporal changes. All of the authors contributed to interpretation and analysis of the data.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.J.S. (Emerson.Speyerer@asu.edu).

**Reviewer Information** Nature thanks M. Cintala, B. Ivanov and the other anonymous reviewer(s) for their contribution to the peer review of this work.



## METHODS

**LROC instrument.** The LROC consists of Wide Angle Camera (WAC) and twin NACs that provide multispectral and high-resolution imaging, respectively. The twin NACs, designated as NAC-Left (NAC-L) and NAC-Right (NAC-R), are capable of acquiring panchromatic images at 0.5 m per pixel from an altitude of 50 km. A typical NAC image consists of 5,064 samples and 52,224 lines, resulting in over 500 megapixels of image data in each NAC-L and -R observation pair. More information on the cameras and images can be found in ref. 31 and at <http://lroc.sese.asu.edu>. As of 15 June 2015, the LROC NACs have acquired over one million images of illuminated terrain. From this image collection, 14,092 are observations of regions of the Moon where previous NAC observations with similar lighting geometry exist (Extended Data Fig. 2). These before and after image pairs, called 'temporal pairs' here, enable the search for changes to the lunar surface that occurred between the times when the first and second images were captured. **NAC temporal imaging.** For this analysis, we define a NAC temporal pair as a set of NAC observations of the same spatial area acquired under similar lighting conditions: an incidence angle of less than 50° to avoid shadows and a difference of less than 3° between the incidence angles of the before and after observations. The temporal gaps between observations span 176 to 1,241 Earth days (mean of 490 days), owing to the orbital geometry of the spacecraft and the Moon's illumination; the ground sampling distance ranged from 0.47 m to 1.75 m (Extended Data Fig. 3). 138 images have been manually scanned<sup>32</sup> and 19 craters and 638 other surface changes were identified. However, manual scanning of the entire collection of temporal pairs is impractical. We systematically scanned more than 2.1 trillion pixels through a semi-automated process to cover more than 6.6% of the Moon. To develop an algorithm to identify surface changes, we used the collection of surface changes identified in ref. 32 as a baseline.

Temporal imaging provides a means to derive a rate of surface change without the concern of the changes saturating the surface. In the case of impact craters, temporal pairs can derive the unsaturated crater-production rate for small craters. For small impacts craters, down to 10 m in diameter, the Neukum production function (NPF) and corresponding chronology function relied on crater statistics for the Cone and North Ray craters<sup>2</sup>. Older surfaces of the Moon are eroded and saturated with craters with diameters of 1 km and smaller, making it impossible to derive an accurate production function in those regions. Therefore, the cratering rates derived from temporal pairs will ultimately help to approximate ages of very young (<50 Myr) features such as irregular mare patches and fresh lobate scarps on the Moon, and young terrain on other planetary bodies, with the aid of scaling laws. Additionally, temporal pairs reveal and monitor other surface changes (that is, splotches and mass wasting events), surface effects of which cannot always be quantified using only post-image analysis.

**Automated change detection.** Change Recognition using Images with Similar Phase (CRISP) is an automated change-detection program developed by the LROC team to identify surface changes using temporal observations. The program is split into three functions: image registration, pixel- and area-based change-detection filtering, and image segmentation. CRISP starts with the Planetary Data System (PDS) formatted and archived experiment data records (EDRs) for each NAC observation that constitutes the temporal pair. Before inspecting the images for changes, we apply a radiometric calibration<sup>31</sup> to each image, which converts the digital numbers (DN) in the EDR image into reflectivity (irradiance/solar flux,  $I/F$ ). We then rectify the images into a common reference frame using the spacecraft ephemeris, a temperature-dependent pointing model for each NAC, and a numerical camera model that defines the internal geometry of the sensor and optics<sup>33</sup>. This procedure enables images to be aligned within 20 m without requiring any ground control points. Although sufficient for many cartographic purposes, a misregistration of 20 m (40 pixels acquired at an altitude of 50 km) is too coarse for small-scale change detection (Extended Data Fig. 4).

To achieve proper registration (<1 pixel), such that each surface feature in the before image is matched to the same feature in the after image, the image pairs are registered using an automatic co-registration tool called *coreg* in the Integrated Software for Imagers and Spectrometers (ISIS) developed by the US Geologic Survey Astrogeology Science Center<sup>34</sup>. *coreg* is an area-based image-matching program that takes a pattern chip from one image and compares it to a similarly sized subregion in a larger search chip. For each pixel in the search chip, *coreg* computes a goodness-of-fit parameter to determine the coordinates of the best image match. Here we use a normalized 2D cross-correlation factor<sup>35</sup> to quantitatively compare the pattern and search window chips.

Once registered, we compute an image ratio by dividing the after image by the before image (Fig. 1a–c). Because the lighting and viewing geometries are similar and the images are registered within a pixel, the values in the ratio image are close to one except in cases where a change has occurred on the surface that affects the reflectance. This reflectance change can be caused by a new impact event, the emplacement of a splotch or a recent mass-wasting event.

However, there are still some variations in the reflectance caused by natural features on the Moon. For example, the temporal ratio image can encounter sensor noise in shadowed portions of the image where the captured signal is low. Additionally, if there are slight differences in lighting or viewing geometries differences in the before and after images, then small photometric differences can cause disparities in the resultant temporal ratio image. To separate the variations due to natural lighting and sensor noise, we create a second change-detection filter using a normalized 2D cross-correlation (NCC) factor for an  $n \times n$  pixel area in the before image compared to the same-sized area in the same location of the after image. The values are similar to those computed using *coreg*, except that the pattern chip does not scan across the search chip. For an area with no detectable surface changes, the corresponding area in the NCC image contains values near one. When the  $n \times n$  pixel NCC filter encounters a substantial texture difference between the before and after images, the recorded values in the NCC deviate from unity.

Together, the ratio image and NCC image are used to identify surface changes in the before and after image pairs. The threshold values were empirically derived using the surface changes identified through manual scanning of 138 NAC temporal pairs<sup>32</sup>. For the change to be detected by the algorithm, the reflectance in one of the pixels covering the feature must change by 2% in the ratio image. For each detected change identified by CRISP, a 200 × 200 pixel cut-out is generated for the before image, the after image, the ratio image and the change mask. These full-resolution sub-images are combined into an animated gif image with metadata associated with the temporal pair (for example, image names, location and temporal difference between observations). Additional metadata and information regarding the animated gif is stored in a PostgreSQL database for continued analysis.

**Change classification.** Detected changes are classified manually with an interface that blinks the sub-images (animated gifs) and displays associated characteristics that the user can flag: low-reflectance change, high-reflectance change, visible crater rim, associated rays and indication of emplacement direction. On the basis of measurements collected during our study, a user can classify a batch of thumbnails on an average of one classification every 3.3 s. These user-supplied classification flags are integrated into the PostgreSQL database for later analysis. This process not only catalogues the surface changes, but also removes any of the false-positives identified by CRISP. For a change to be classified as a splotch or new impact crater, it must be unambiguous. Therefore, we omit cataloguing some of the very small impacts (fewer than four pixels) and small splotches (fewer than about five pixels), owing to a lack of visual evidence in the temporal image pair. Therefore, we are underestimating the population of small surface changes, and surface changes that are not creating distinct contrast boundaries with their surroundings.

In cases where a rim is identified, we measure and record the diameter of the crater. For an accurate diameter measurement, the crater must span at least four pixels. For our temporal dataset, for which the ground sampling distance ranged from 0.47 m to 1.75 m (Extended Data Fig. 3), the crater with the smallest diameter that could be visible in all temporal pairs is 7 m across (1.75 m per pixel × 4 pixels). **Normalizing spatial coverage with respect to time.** Because the current spacing between NAC temporal-pair observations varies between 176 and 1,241 days, we scaled the search area for each NAC temporal-pair observation by calculating the product of the actual area searched ( $A_i$ ) and the temporal gap ( $\Delta t_i$ ), and dividing it by a year period to maintain the units of area. The sum of this expression results in the total annual search area:

$$A_{\text{annual}} = \sum_{i=1}^n \frac{A_i \Delta t_i}{1 \text{ yr}}$$

This technique, which was used in ref. 36 to derive the contemporary cratering rate on Mars, yielded an annual surface area of  $3.31 \times 10^6 \text{ km}^2$ , or 8.71% of the Moon observed by NAC temporal imaging (original non-scaled area of  $2.49 \times 10^6 \text{ km}^2$  or 6.57% of the Moon). Using this normalized surface area, we derived cratering rates (Fig. 2) and a production function for splotches (Fig. 4e).

**Cratering rate.** Using the crater statistics and the normalized surface area ( $3.305 \times 10^6 \text{ km}^2 \text{ yr}^{-1}$ ), we evaluated the current crater-production rate (Fig. 2 and Extended Data Fig. 5). To provide a basis for comparison, we evaluated our statistics with the NPF, which is a standard used to estimate model ages for geologic units using size–frequency statistics of craters with diameters between 10 m and 300 km (ref. 2). The NPF predicts that 140 craters with diameters of  $\geq 10$  m form annually on the Moon. From the area covered by the NAC temporal pairs scaled to a 1-yr period, the NPF estimates that 12 craters ( $D \geq 10$  m) should have been located. However, the 14,092 randomly targeted temporal pairs revealed 16 impact craters ( $D \geq 10$  m). This total excludes the 18-m and 34-m craters discovered by targeted observation after observed impact flashes<sup>7,9</sup>. Using a hypergeometric distribution that accounts for the area searched and the modelled crater-production rate, we found that there is only a 16% probability of finding 16 or more impact

craters in the NAC temporal dataset. Therefore, on the basis of our tentative crater statistics, the current crater-production rate is potentially higher than the NPF predicts.

To see the effect of an increased cratering rate on an absolute model age, we fit the NPF to the 15 new impact craters discovered with diameters of 10–20 m and found an age that is 20% older than modelled when assessed with a Poisson timing analysis<sup>10</sup> (measured age of  $1.2 \pm 0.3$  yr). In this initial analysis, we omitted the largest crater (43 m) from our fitting, because craters in that size range are not common enough to form on an annual basis in the search area. The NPF predicts that 1–2 (1.56) craters with diameters of  $\geq 43$  m form annually on the entire Moon<sup>2</sup>. Given the search area, there is only a 14% chance of finding a crater with a diameter of  $\geq 43$  m. Therefore, including the 43-m crater in our absolute model age derivation could falsely increase the model age. However, if we included the 43-m crater, then the measured model age increases to  $1.3 \pm 0.3$  yr. As LROC acquires more temporal-pair observations throughout the rest of the extended mission, the increased spatial and crater frequency statistics collected will enable us to refine the size–frequency distribution of craters with diameters of up to 75–100 m.

Using the same crater population ( $10 \text{ m} \leq D \leq 20 \text{ m}$ ), we found that the cumulative power-law slope ( $-b$ ) derived without binning the craters is potentially steeper than the NPF estimates ( $b = 4.62$ ; 95% CI of 4.15–5.14 and 2.97–3.15, respectively). A fit to the differential size–frequency distribution of the craters also yielded a potentially steeper slope compared to the NPF ( $b_{\text{diff}} = 5.11$ ; 95% CI of 2.27–6.94 and 3.97–4.15, respectively). These findings imply that if the crater-production rate is higher in this diameter range, then the model ages derived with the NPF using small craters near the 10-m cut-off may contain a systematic offset.

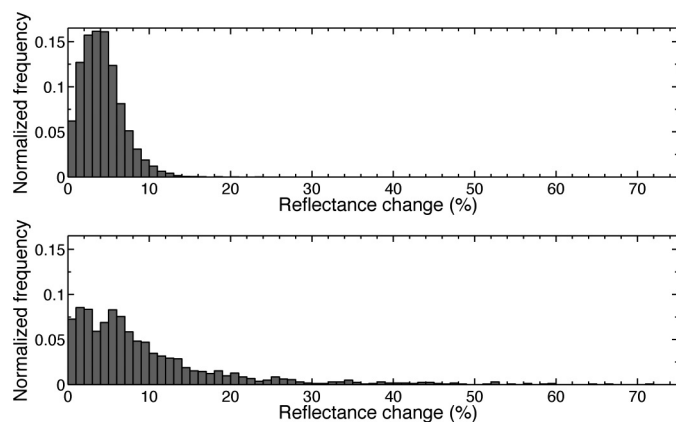
In addition to the NPF, there are other crater production functions derived from our current knowledge of the asteroid and comet population in the inner Solar System. First, the model production function (MPF) in ref. 37 with the Holsapple and Housen scaling laws applied predicts fewer craters forming in the size range examined here and a distribution with a similar cumulative slope to that of the NPF. Second, the crater size–frequency distribution in ref. 11 that is valid for crater diameters between 100 m and 1,000 km predicts that we should have located 17 craters when we extrapolated this distribution down to craters with diameters of  $\geq 10$  m (we identified 16). Additionally, for the 10–20-m diameter range, the cumulative power-law slope ranges between 3.56 and 4.50 in the extrapolated model<sup>11</sup>, which overlaps the confidence interval we derived from our dataset (4.62; 95% CI of 4.15–5.14). However, this increased rate and steeper slope seen in the extrapolated model and data presented in ref. 11 contradicts other models based on counts of terrestrial bolides<sup>38</sup> and observations of shallower power-law slopes derived from new terrestrial bolides<sup>39</sup> and recent Martian craters<sup>36</sup>. However, this shallower slope in the Martian dataset might be the result of rapid weathering of the blast zones observed around small Martian craters, which could make it difficult to identify the smallest new craters<sup>40</sup>. Analysis of additional temporal image pairs will ultimately strengthen our crater statistics and will help to refine the rate of crater production and to validate cratering models.

**Characterizing impact sites.** We found that each new impact crater discovered over the course of this investigation was surrounded by up to four distinct reflectance zones that are the result of the crater-formation process. One of the clearest examples of all four reflectance zones is around the impact crater that formed on 17 March 2013 (ref. 7) (Extended Data Fig. 6). We identified the visible zones and

measured the maximum distance spanned by each reflectance zone from the centre of each new crater (Extended Data Fig. 7 and Extended Data Table 1). Statistics regarding the span of the proximal zones are provided in Extended Data Table 2. Owing to the difficulty in measuring the size of small craters that are near the resolution limits of the observations, we omitted craters smaller than 3 m from the reflectance-zone analysis; the statistics regarding the 43-m crater are omitted in the figures, owing to a lack of adequate before imaging near the impact site.

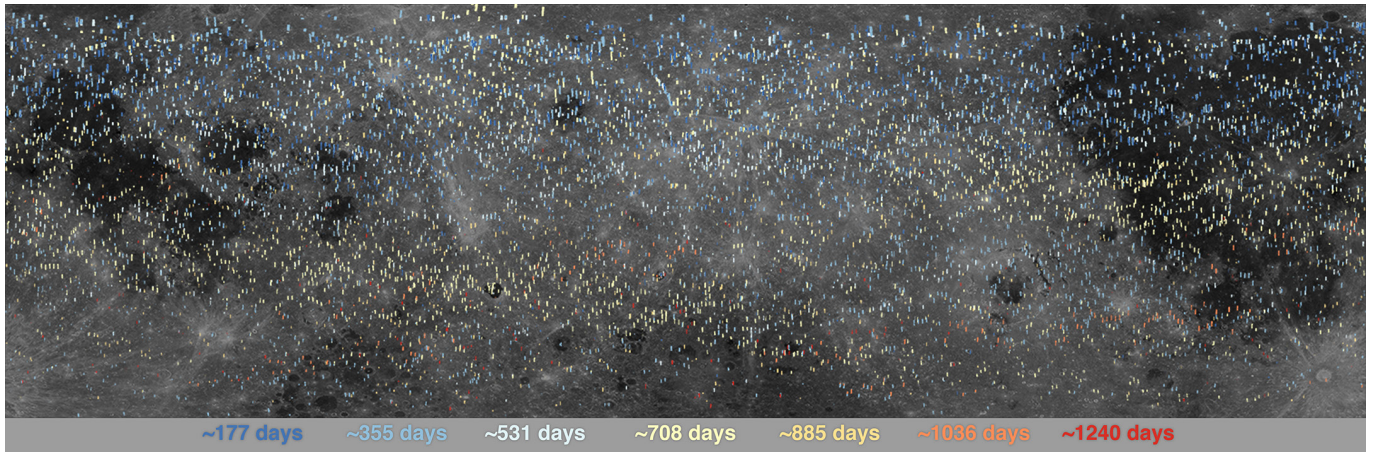
**Modelling splotch accumulation.** The varying spatial resolution of the temporal pairs causes a ‘roll-off’ in the size–frequency distribution for splotches with diameters of  $< 10$  m (Fig. 4e). Using a least-squares power-law fit for splotches with diameters of 10 m (the diameter at which splotches are resolved at all pixel scales) to 40 m, we estimate that  $0.0029 \pm 0.0001$  splotches with diameters of  $> 10$  m form over an area of  $1 \text{ km}^2$  annually (Fig. 4e), which is equivalent to  $1.09 \times 10^5$  splotches forming across the entire Moon annually. Using the power-law fit ( $n = ad^{-b}$ ;  $a = 1.483 \times 10^{-11}$ , 95% CI of  $(1.459\text{--}1.508) \times 10^{-11}$ ;  $b = 4.144$ , 95% CI of  $4.148\text{--}4.141$ ;  $d$  is the splotch diameter in metres;  $R^2 = 0.9991$ ) and extrapolating the results to 5 m and to 1 m yields an annual rate of  $0.0510 \pm 0.0018$  and  $40.2 \pm 1.6$  splotches per square kilometre, respectively, with an annual global accumulation of  $1.93 \times 10^6$  and  $1.52 \times 10^9$ , respectively. Using this splotch production function and a Monte Carlo model that randomly emplaces splotches, we estimate that 99% of the lunar surface would be altered after  $1.0 \times 10^7$  yr from splotches with diameters of  $\geq 10$  m. By combining the coverage rate and assuming a conservative churn depth with respect to splotch diameter, we can estimate regolith gardening<sup>5,13</sup> rates from the emplacement of splotches on the surface. Using a ratio of depth to diameter of 1:50 for these features, which is a conservative estimate based on laboratory experiments of clustered impacts at similar speeds<sup>25</sup>, we estimate that these events would effectively churn the upper 20 cm of regolith after  $1.0 \times 10^7$  yrs. Extrapolating the rate of splotch formation down to 5 m and 1 m drastically reduces the timescales to  $2.3 \times 10^6$  yr and  $8.1 \times 10^4$  yr, respectively. Applying the same assumption regarding the ratio of depth to diameter to the latter simulation yields a regolith gardening rate of about  $10^3$  yr for the upper 2 cm.

31. Robinson, M. S. *et al.* Lunar Reconnaissance Orbiter Camera (LROC) instrument overview. *Space Sci. Rev.* **150**, 81–124 (2010).
32. Thompson, S. D., Bowles, Z. R., Povilaitis, R. Z., Daubar, I. J. & Robinson, M. S. Recent impacts on the Moon. *45th Lunar and Planet. Sci. Conf. abstr.* 2769 (2014).
33. Speyerer, E. J. J. *et al.* Pre-flight and on-orbit geometric calibration of the lunar reconnaissance orbiter camera. *Space Sci. Rev.* **200**, 357–392 (2016).
34. Anderson, J. A., Sides, S. C., Soltesz, D. L., Sucharski, T. L. & Becker, K. J. Modernization of the Integrated Software for Imagers and Spectrometers. *Lunar Planet. Sci. Conf.* **35**, abstr. 2039 (2004).
35. Gonzalez, R. C. & Woods, R. E. *Digital Image Processing* (Addison-Wesley, 1992).
36. Daubar, I. J., McEwen, A. S., Byrne, S., Kennedy, M. R. & Ivanov, B. The current martian cratering rate. *Icarus* **225**, 506–516 (2013).
37. Marchi, S., Mottola, S., Cremonese, G., Massironi, M. & Martellato, E. A new chronology for the Moon and Mercury. *Astron. J.* **137**, 4936–4948 (2009).
38. Ivanov, B. A. Earth/Moon impact rate comparison: searching constraints for lunar secondary/primary cratering proportion. *Icarus* **183**, 504–507 (2006).
39. Brown, P. G. *et al.* A 500-kiloton airburst over Chelyabinsk and an enhanced hazard from small impactors. *Nature* **503**, 238–241 (2013).
40. Daubar, I. J. *et al.* Changes in blast zone albedo patterns around new martian impact craters. *Icarus* **267**, 86–105 (2016).

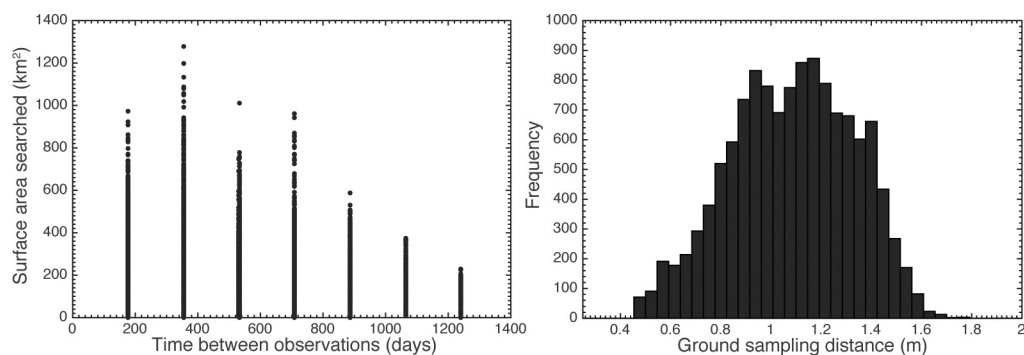


**Extended Data Figure 1 | Reflectance changes caused by emplacement of splotches.** Histogram of reflectance changes associated with low-reflectance (top;  $n = 18,756$ ) and high-reflectance (bottom;  $n = 1,757$ ) splotches.

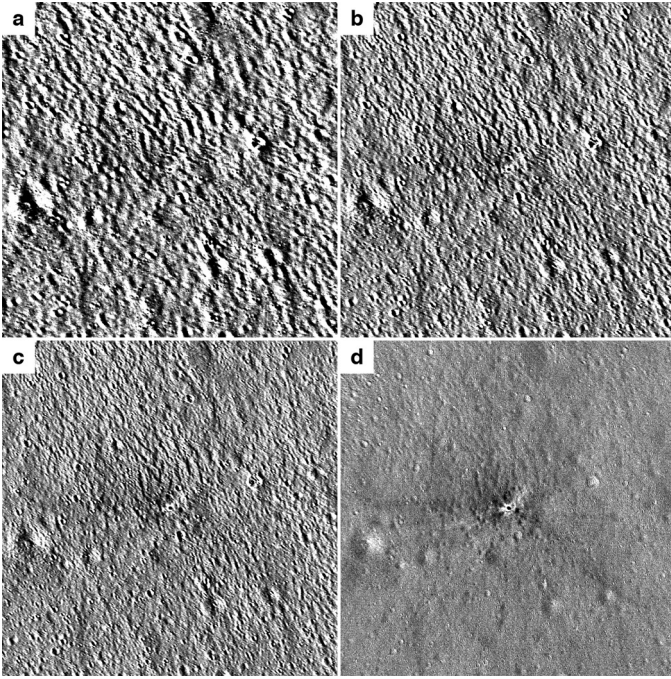




**Extended Data Figure 2 | Distribution of NAC temporal pairs.** Image footprints colour-coded by the number of days between the before and after observations (June 2009 to May 2015). The size of image footprints are exaggerated for display clarity.



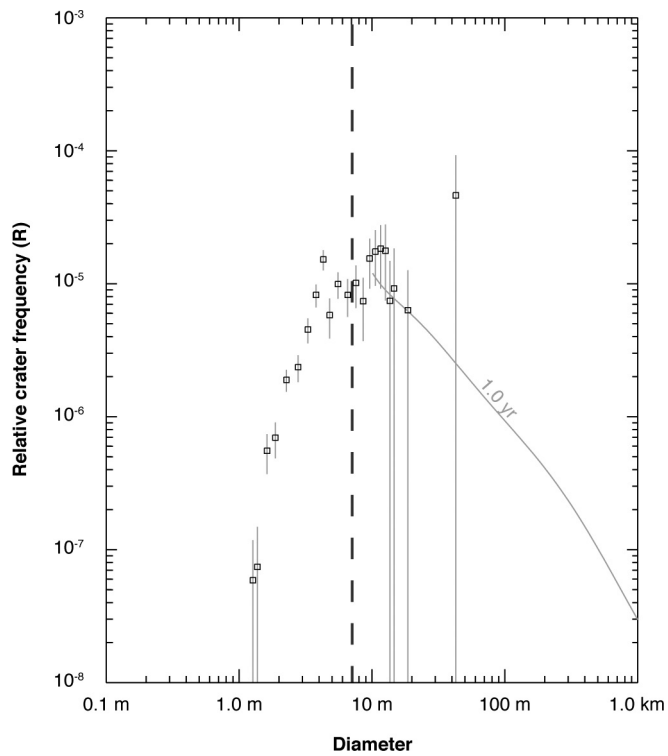
**Extended Data Figure 3 | Temporal-pair statistics.** Left, distribution of the area covered by individual temporal pairs versus the temporal spacing between the corresponding observations. Right, histogram of the ground sampling distance for each after image of the NAC temporal pair.



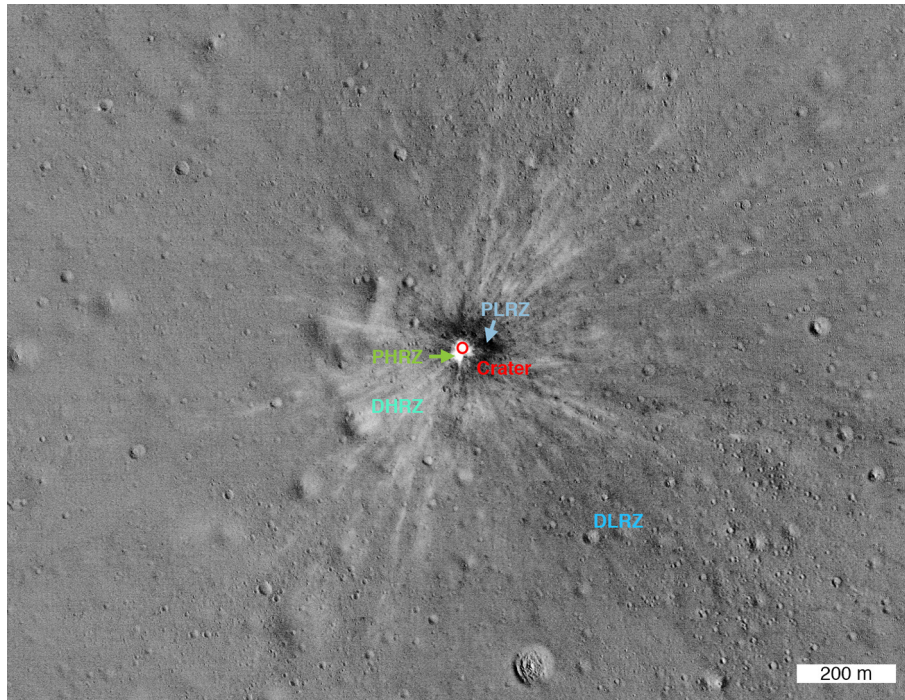
**Extended Data Figure 4 | Effect of image registration errors.**

**a–d**, Examples of a temporal ratio image with decreasing pixel offsets (ratio of NAC frames M188678240LR/M1180548227LR): 10-pixel offset (7.8-m offset; **a**); 5-pixel offset (3.9-m offset; **b**); 3-pixel offset (2.3-m offset; **c**); and <1-pixel offset (<0.8-m offset; **d**). The larger offsets in **a–c** make the identification of the new impact crater impossible.

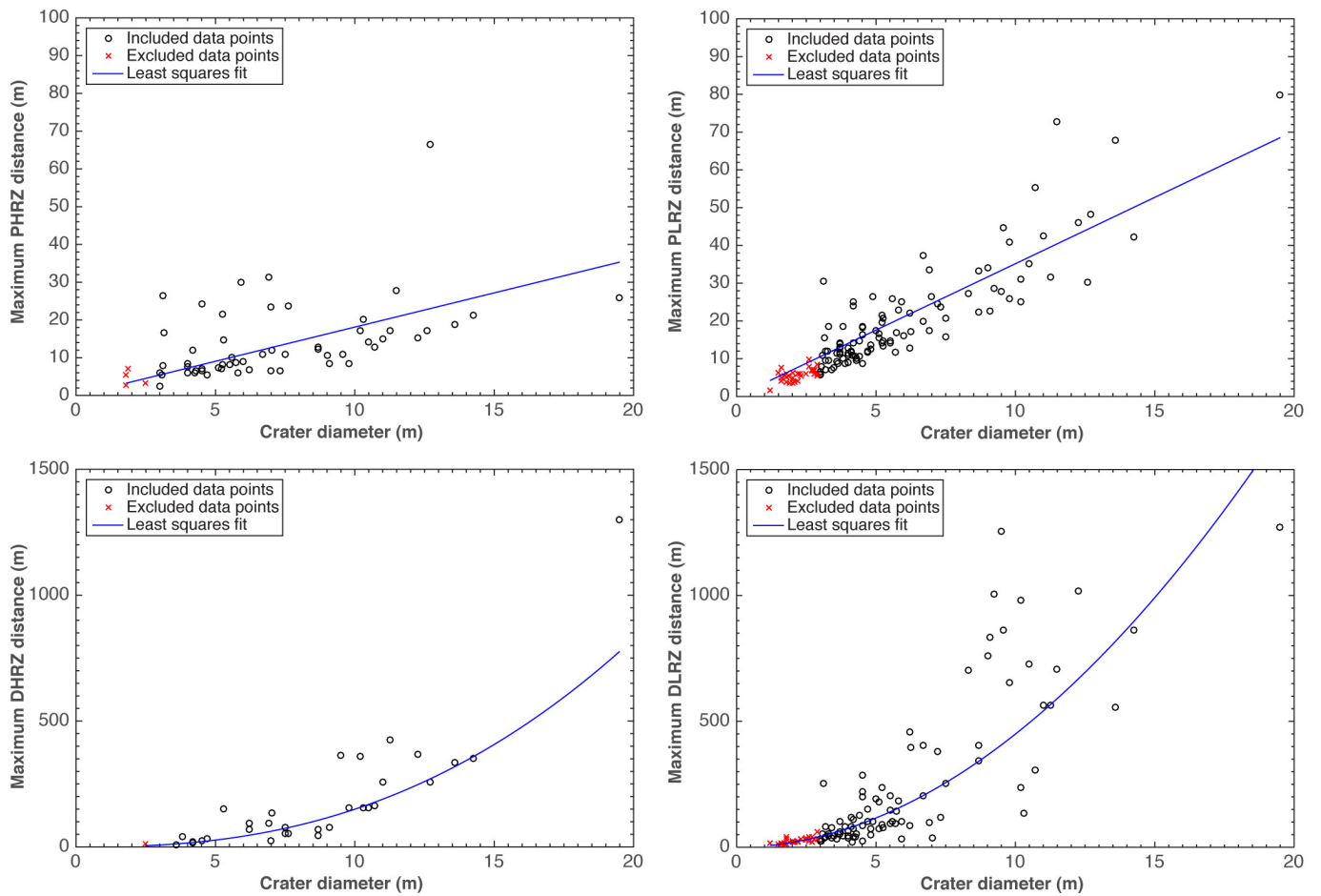




**Extended Data Figure 5 | R-plot of new crater population.** Relative crater frequency  $R$  of the 222 new impact craters identified with temporal imaging. For reference, a 1-yr isochron (grey line) derived from the NPF is overlaid for diameters of  $\geq 10$  m. Error bars are estimated on the basis of Poisson statistics of counts.



**Extended Data Figure 6 | Details of the impact process recorded in the temporal pair.** Temporal ratio image of the 17 March 2013 impact site surrounded by four distinct reflectance zones (ratio of NAC frames M1129645568L/M183689789L).



**Extended Data Figure 7 | Range of reflectance zones associated with new impacts.** Maximum zone distance versus crater diameter for each of the four reflectance zones observed around new impacts. Craters smaller than 3 m in diameter were excluded from the least-squares fit.



**Extended Data Table 1 | Least-squares fit to maximum zone distance compared to crater diameter**

Zone	Fit type	a (95% CI)	b (95% CI)	R <sup>2</sup>
PHRZ	Linear ( $aD+b$ )	1.81 (1.50 to 2.11)	0 (fixed)	0.169
PLRZ	Linear ( $aD+b$ )	3.52 (3.32 to 3.71)	0 (fixed)	0.747
DHRZ	Power ( $aD^b$ )	0.52 (0.11 to 0.92)	2.46 (2.17 to 2.75)	0.963
DLRZ	Power ( $aD^b$ )	5.03 (3.84 to 6.21)	1.95 (1.86 to 2.04)	0.981

Fit information for the plots shown in Extended Data Fig. 7, where  $D$  is the diameter of the crater in metres.

**Extended Data Table 2 | Span of the proximal reflectance zones**

Zone	Mean	Median	Standard deviation	Minimum distance	Maximum distance
PHRZ	$2.11D$	$1.50D$	$1.46D$	$0.84D$	$8.56D$
PLRZ	$3.16D$	$2.91D$	$1.11D$	$1.33D$	$9.85D$

Statistical summary of the span of each proximal reflectance zone. Each value has been normalized to the crater diameter  $D$ .

# Enhanced flexoelectric-like response in oxide semiconductors

Jackeline Narvaez<sup>1</sup>, Fabian Vasquez-Sancho<sup>1,2</sup> & Gustau Catalan<sup>1,3</sup>

**Flexoelectricity is a property of all dielectric materials whereby they polarize in response to deformation gradients such as those produced by bending<sup>1–5</sup>. Although it is generally thought of as a property of dielectric insulators, insulation is not a formal requirement: in principle, semiconductors can also redistribute their free charge in response to strain gradients. Here we show that bending a semiconductor not only generates a flexoelectric-like response, but that this response can in fact be much larger than in insulators. By doping single crystals of wide-bandgap oxides to increase their conductivity, their effective flexoelectric coefficient was increased by orders of magnitude. This large response can be explained by a barrier-layer mechanism that remains important even at the macroscale, where conventional (insulator) flexoelectricity otherwise tends to be small. Our results open up the possibility of using semiconductors as active ingredients in electromechanical transducer applications.**

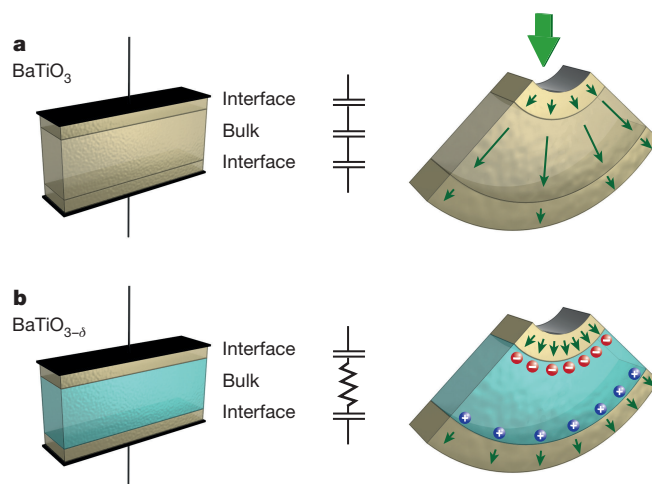
Unlike piezoelectricity (polarization induced by homogeneous deformations such as stretching or compressing), flexoelectricity (polarization induced by inhomogeneous strain such as bending) is allowed by symmetry in all materials and is therefore a more widespread property of solids. However, the charge density that can be generated by flexoelectricity is normally small, which limits its practical appeal. This has motivated a search for ways to enhance flexoelectricity, which has so far yielded two tried-and-tested strategies. The first involves exploiting the proportionality between flexoelectricity and permittivity<sup>1–5</sup>: materials with high dielectric constants also have high flexoelectric coefficients<sup>2,6</sup>. The second involves maximizing strain gradients by working at very small (nanoscopic) size scales<sup>7–9</sup>, at which maximum achievable deformations are larger—put simply, a thin film can be bent more than a thick slab. The first strategy forces one to work with high-permittivity materials and the second limits the size range in which flexoelectricity is competitive. Achieving larger responses in a wider range of materials and, crucially, to do so at the macroscale, are among the most pressing challenges in the field.

Although bulk flexoelectricity has a theoretical limit that cannot be exceeded<sup>1,3,5</sup>, there are caveats to this limitation. First, bending a material not only elicits polarization from bulk flexoelectricity, but also from surface piezoelectricity<sup>3,5,10–13</sup>, which is caused by the strains on the opposite sides of a bent crystal (compression on the concave side, extension on the convex side). Because all surfaces are asymmetric (and therefore piezoelectric), surface piezoelectricity is, from the points of view of symmetry and functionality, as universal as bulk flexoelectricity, and is in fact regarded as an actual intrinsic component of the total flexoelectric response<sup>5,10–13</sup>.

The second caveat has to do with the permittivity itself, which is a measure of the polarizability of a material. In a dielectric insulator, it is given by how much (and how easily) positive and negative bound charges can be separated. However, the effective polarizability of a capacitor structure (that is, its capacitance) can be enhanced above the dielectric limit if we also allow free charges to separate. The physics of

this phenomenon is described by the Maxwell–Wagner model<sup>9</sup> and is exploited within the capacitor industry to increase the storage density of so-called ‘barrier-layer capacitors’<sup>14,15</sup>. The basic idea is depicted in Fig. 1b: when an electric field is applied to a heterogeneous material consisting of insulating barrier layers separated by a (semi)conducting region, the conducting region responds by allowing its free charges to move across the barrier layers, thus effectively behaving as an intercalated electrode. In this scenario, only the thin barrier layers contribute to the capacitance and, because capacitance is inversely proportional to thickness, enormous capacitances<sup>14–18</sup> and even increased piezoelectricity<sup>19</sup> can be achieved. As we show here, the barrier-layer mechanism can elicit much more bending-induced charge from semiconductors than from insulators.

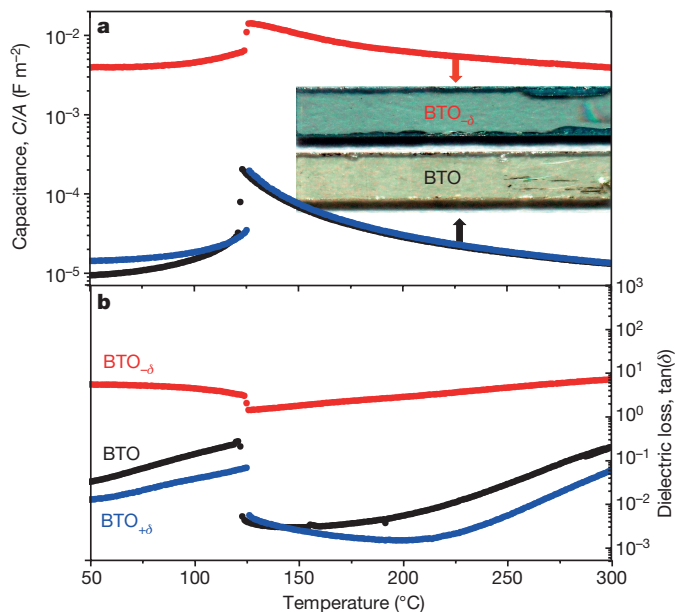
To explore this idea, we studied single crystals of BaTiO<sub>3</sub> doped with oxygen vacancies to increase their conductivity<sup>20,21</sup>. Henceforth, as-received BaTiO<sub>3</sub> is labelled BTO; oxygen-depleted BaTiO<sub>3–δ</sub> (made by vacuum annealing, see Methods) is labelled BTO<sub>–δ</sub>, and subsequently re-oxidized BaTiO<sub>3</sub> made by re-annealing in an oxygen atmosphere (see Methods) is referred to as BTO<sub>+δ</sub>. Fully oxidized BTO is an



**Figure 1 | Barrier layer model.** **a**, A dielectric responds to voltage as three capacitors in series, with one bulk and two interfacial layers. The same is true for its electromechanical response to bending, with polar contributions (sketched as green arrows) from bulk flexoelectricity and surface piezoelectricity. **b**, A semiconductor can form depletion layers at the interfaces with the electrodes; in these circumstances, the conducting bulk acts as an intercalated electrode (blue layer) between interfacial barrier layers that respond as thin capacitors. This results in larger capacitance (inversely proportional to barrier thickness) and enhanced surface piezoelectricity, owing to screening of the internal depolarizing field by free charges (sketched as blue and red spheres).

<sup>1</sup>Institut Català de Nanociència i Nanotecnologia (ICN2), CSIC and The Barcelona Institute of Nanoscience and Technology (BIST), Campus UAB, 08193 Barcelona, Spain. <sup>2</sup>Centro de Investigación en Ciencia e Ingeniería de Materiales, Universidad de Costa Rica, San José 11501, Costa Rica. <sup>3</sup>ICREA—Instituto Catalana de Recerca i Estudis Avançats, Passeig Lluís Companys 23, 08010 Barcelona, Spain.



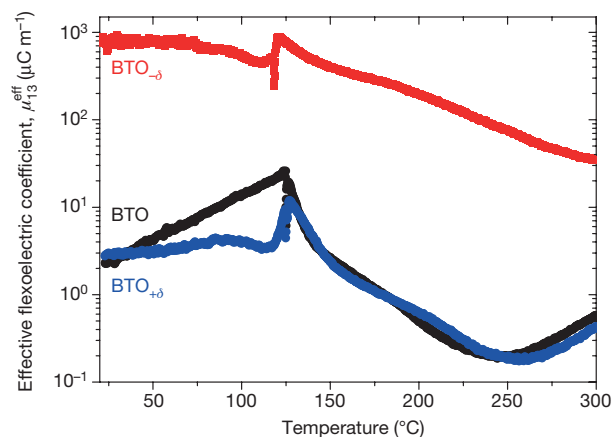


**Figure 2 | Capacitance of BTO.** **a**, **b**, Capacitance per unit area,  $C/A$ , (**a**) and dielectric loss tangent (**b**) as a function of temperature for BTO (black),  $\text{BTO}_{-\delta}$  (red) and  $\text{BTO}_{+\delta}$  (blue). The inset in **a** shows plan-view photographs of the reduced (top) and oxidized (bottom) crystals. The reduced crystal has a darker colour owing to increased light absorption by free carriers (the even darker regions at the edges are shadow effects from the illumination; the edges become chipped as a result of thermal stress during annealing).

archetypal ferroelectric with high permittivity and flexoelectricity<sup>4–6</sup>. In terms of conductivity, it is a wide-bandgap (3.5 eV) material with reasonably good dielectric insulation. However, when doped with oxygen vacancies, BTO becomes an n-type electronic semiconductor<sup>20,21</sup> with charge-depleted surfaces<sup>21,22</sup>.  $\text{BTO}_{-\delta}$  is also a good ionic conductor at high temperature<sup>23</sup>, with the vacancies acting simultaneously as electron donors and as ionic charge carriers. This is relevant because ionic defects such as vacancies can also respond to strain gradients via the Vegard effect; such a ‘flexoionic’ effect would be the inverse of the electrochemical strain observed in ionomers, which can be quantitatively comparable to bulk flexoelectricity in mixed ionic–electronic conductors<sup>24</sup>. However, at the relatively low temperatures (by ionic standards) of the present work, the ionic contribution to the conductivity was found to be rather small (see Methods); nevertheless, to exclude the role of ions, we have also studied the response of pure and Nb-doped  $\text{TiO}_2$ , the latter being a vacancy-free electronic semiconductor at room temperature<sup>25</sup>.

The capacitance density and dielectric loss of the BTO crystals are shown in Fig. 2.  $\text{BTO}_{-\delta}$  shows classic signatures of Maxwell–Wagner behaviour: high dissipation (loss tangent), consistent with increased conductivity, and increased capacitance, consistent with the existence of thin insulating barrier layers<sup>9,14,15,17,18,20</sup>. Because the ratio between the capacitance of  $\text{BTO}_{-\delta}$  and BTO is proportional to the thickness ratio between the bulk and the interfacial barrier layers, we can estimate an upper bound of about 1  $\mu\text{m}$  for the thickness of the barrier layer (see Methods).

The effective flexoelectric coefficient  $\mu_{\text{eff}}$  is defined as the change in polarization (measured as charge per unit area collected at the electrodes) divided by applied strain gradient (beam curvature)<sup>5</sup>, and is plotted as a function of temperature in Fig. 3; the measurement set-up (see Methods) is as previously used for other insulating crystals<sup>26</sup>. The results show that the effective flexoelectricity of  $\text{BTO}_{-\delta}$  is two orders of magnitude larger than for insulating BTO. This enhancement persists at temperatures above which the ferroelectric phase (where macroscopic piezoelectricity can yield spurious enhancements<sup>27</sup>) is stable, and also above those at which polar nanoregions exist<sup>26,28</sup>, so it cannot be attributed to residual



**Figure 3 | Effective flexoelectricity of BTO.** The transverse effective flexoelectric coefficient ( $\mu_{13}^{\text{eff}}$ ; charge density collected at the electrodes divided by applied strain gradient) is plotted as a function of temperature for reduced (conducting)  $\text{BTO}_{-\delta}$  and oxidized (insulating) BTO and  $\text{BTO}_{+\delta}$ . The conducting sample shows a two order of magnitude enhancement, reaching into the millicoulomb per metre range.

ferroelectricity either. The intrinsic properties of the as-received (BTO) sample can be recovered upon reoxidation ( $\text{BTO}_{+\delta}$ ), so the enhancement is reversible and clearly caused by the doping.

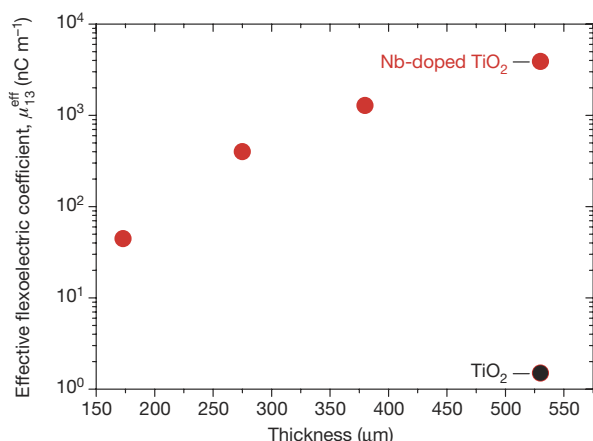
The effective flexoelectric coefficient of  $\text{BTO}_{-\delta}$  reaches values close to  $1 \text{ mC m}^{-1}$ —the largest so far reported for any material<sup>5</sup>. We argue that the origin of this large response is not bulk flexoelectricity, but its close relative, surface piezoelectricity. As mentioned earlier, in an insulator, these two effects are regarded as inseparable components of the total response<sup>3,5,10–13</sup>. However, in a semiconductor with insulating barrier layers such as that depicted in Fig. 1, there are additional considerations: (I) the bulk polarization can be screened by free charges<sup>22</sup>, so bulk flexoelectricity should no longer contribute to the total polarization of the capacitor structure, and this allows us to separate out the surface contribution; and (II) bulk free charges can also screen the depolarization field that would otherwise appear at the interface between the barrier layer and the bulk.

In effect, such a system should behave as a bimorph transducer consisting of two piezoelectric barrier layers attached to the opposite surfaces of a conducting slab that acts as an intercalated electrode. The phenomenology is somewhat reminiscent of Haertling’s ‘rainbow’ transducers<sup>29</sup>, which use asymmetric annealing to turn one side of a piezoelectric material into an electrode. Our crystals are instead symmetric, but there are more fundamental differences: (I) although we have annealed BTO to make it semiconducting, the barrier-layer mechanism can, in principle, appear in any semiconductor (doped or intrinsic) as long as it has high-resistivity interfaces such as Schottky barriers; and (II) the starting material does not need to be piezoelectric because the polarization comes from the surface, which is always piezoelectric. These two points are illustrated by additional measurements performed on the centrosymmetric semiconductor Nb-doped  $\text{TiO}_2$  (rutile), which displays a 1,000-fold flexoelectric enhancement compared to undoped, insulating  $\text{TiO}_2$  (Fig. 4).

In the proposed scenario, the bending-induced charge density is the product of the surface’s transverse piezoelectric constant ( $e_{13}^{\text{surf}}$ ) and the surface strain, while the surface strain is the product of the curvature ( $G$ ) and the half-thickness of the crystal ( $t/2$ ). Hence, the bending-induced surface polarization is  $P = e_{13}^{\text{surf}} G t/2$  and the effective flexoelectric coefficient (defined as polarization  $P$  divided by curvature  $G$ ) is

$$\mu_{13}^{\text{eff}} = e_{13}^{\text{surf}} \frac{t}{2} \quad (1)$$

It is also possible to arrive at this equation by starting from the general expression for the surface-piezoelectric contribution to the effective flexoelectricity of a dielectric<sup>10</sup>



**Figure 4 | Effective flexoelectricity of TiO<sub>2</sub> and 0.05%Nb-doped TiO<sub>2</sub>.** For two crystals of identical thickness, bending the semiconductor (Nb-doped TiO<sub>2</sub>; filled red circles) generates three orders of magnitude more change than does bending the insulator (TiO<sub>2</sub>; black filled circle). The effective flexoelectric coefficient ( $\mu_{13}^{\text{eff}}$ ) of the semiconductor is proportional to sample thickness, consistent with a barrier-layer model.

$$\mu_{13}^{\text{eff}} = e_{13} t_i \frac{t_b \varepsilon_b}{2 t_i \varepsilon_b + t_b \varepsilon_i}$$

(where  $t_i$ ,  $t_b$  are the thicknesses of the interfacial and bulk layers, respectively, and  $\varepsilon_i$ ,  $\varepsilon_b$  are their dielectric constants), assuming  $t_b \gg t_i$  and taking the limit as the bulk dielectric constant tends to infinity, which is physically equivalent to stating that the bulk can perfectly screen any depolarizing field. The role of our conductive bulk thus becomes clear: it not only isolates the contribution from surface piezoelectricity (itself a hitherto unresolved problem<sup>10–13,26</sup>), but it also enhances this contribution by screening the polar discontinuity at the interface between the piezoelectric barrier layer and the bulk, thus allowing the surface to reach polarization values that would be unattainable in a finite-permittivity insulator. We can also use Equation (1) to calculate the size of the surface piezoelectric coefficient (see Methods). For BTO<sub>1–δ</sub>, we found the surface's transverse piezoelectric charge coefficient (piezoelectric constant  $e$  multiplied by elastic compliance),  $d_{13}^{\text{surf}}$ , to range from 37 pC N<sup>–1</sup> at the Curie temperature peak,  $T_C = 125^\circ\text{C}$ , to 0.6 pC N<sup>–1</sup> (similar to the coefficient of quartz) at 300°C; for comparison, the transverse piezoelectric coefficient of BaTiO<sub>3</sub> at room temperature is 36.5 pC N<sup>–1</sup> (ref. 30). The piezoelectric coefficients of surface layers are therefore comparable to the bulk piezoelectric coefficients of standard piezoelectrics and do not need to be 'giant' to yield a large flexoelectric-like response.

According the barrier-layer model, and as shown by the results in Fig. 4, the effective flexoelectric coefficient of semiconductors is proportional to the thickness, so it cannot be considered a material constant. Instead, the material constant that dictates the flexoelectric-like response of the device is the surface piezoelectric coefficient. In macroscopically thick samples, this surface contribution can not only quantitatively match bulk flexoelectricity, as predicted for insulators<sup>10,12</sup>, but for semiconductors it can surpass it. Thus, the thickness proportionality of the effective flexoelectric coefficient means that the electromechanical response is large not only at the nanoscale, at which conventional (insulator) flexoelectricity is already important, but also at the macroscale. The magnitude of the effect, its persistence at the macroscale, and the ubiquity (and, generally, lead-free composition) of semiconducting materials are all positive news for device design and will hopefully stimulate new research into semiconductor-based flexoelectric and electromechanical applications.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 March; accepted 28 July 2016.

Published online 26 September 2016.

- Kogan, S. M. Piezoelectric effect during inhomogeneous deformation and acoustic scattering of carriers in crystals. *Sov. Phys. Solid State* **5**, 2069–2070 (1964).
- Bursian, E. & Zaikovskii, O. I. Changes in curvature of ferroelectric film due to polarization. *Sov. Phys. Solid State* **10**, 1121 (1968).
- Tagantsev, A. K. Piezoelectricity and flexoelectricity in crystalline dielectrics. *Phys. Rev. B* **34**, 5883–5889 (1986).
- Cross, L. E. Flexoelectric effects: charge separation in insulating solids subjected to elastic strain gradients. *J. Mater. Sci.* **41**, 53–63 (2006).
- Zubko, P., Catalan, G. & Tagantsev, A. K. Flexoelectric effect in solids. *Annu. Rev. Mater. Res.* **43**, 387–421 (2013).
- Ma, W. & Cross, L. E. Flexoelectricity of barium titanate. *Appl. Phys. Lett.* **88**, 232902 (2006).
- Catalan, G. *et al.* Flexoelectric rotation of polarization in ferroelectric thin films. *Nat. Mater.* **10**, 963–967 (2011).
- Lee, D. *et al.* Giant flexoelectric effect in ferroelectric epitaxial thin films. *Phys. Rev. Lett.* **107**, 057602 (2011).
- Lu, H. *et al.* Mechanical writing of ferroelectric polarization. *Science* **336**, 59–61 (2012).
- Tagantsev, A. K. & Yurkov, A. S. Flexoelectric effect in finite samples. *J. Appl. Phys.* **112**, 044103 (2012).
- Stengel, M. Microscopic response to inhomogeneous deformations in curvilinear coordinates. *Nat. Commun.* **4**, 2693 (2013).
- Stengel, M. Surface control of flexoelectricity. *Phys. Rev. B* **90**, 201112 (2014).
- Hong, J. & Vanderbilt, D. First-principles theory of frozen-ion flexoelectricity. *Phys. Rev. B* **84**, 180101 (2011).
- Sinclair, D. C., Adams, T. B., Morrison, F. D. & West, A. R. CaCu<sub>3</sub>Ti<sub>4</sub>O<sub>12</sub>: one-step internal barrier layer capacitor. *Appl. Phys. Lett.* **80**, 2153–2155 (2002).
- Glaister, R. M. Barrier-layer dielectrics. *Proc. IEE Part B* **109**, 423–431 (1962).
- Von Hippel, A. *Dielectrics and Waves* (Artech House, 1995).
- O'Neill, D., Bowman, R. M. & Gregg, J. M. Dielectric enhancement and Maxwell-Wagner effects in ferroelectric superlattice structures. *Appl. Phys. Lett.* **77**, 1520–1522 (2000).
- Catalan, G. & Scott, J. F. Magnetoelectrics: is CdCr<sub>2</sub>S<sub>4</sub> a multiferroic relaxor? *Nature* **448**, E4–E5 (2007).
- Damjanovic, D., Demartin Maeder, M., Duran Martin, P., Voisard, C. & Setter, N. Maxwell-Wagner piezoelectric relaxation in ferroelectric heterostructures. *J. Appl. Phys.* **90**, 5708–5712 (2001).
- Kolodiazny, T. *et al.* Thermoelectric power, Hall effect, and mobility of n-type BaTiO<sub>3</sub>. *Phys. Rev. B* **68**, 085205 (2003).
- Heywang, W. Semiconducting barium titanate. *J. Mater. Sci.* **6**, 1214–1226 (1971).
- Genenko, Y. A., Hirsch, O. & Erhart, P. Surface potential at a ferroelectric grain due to asymmetric screening of depolarization fields. *J. Appl. Phys.* **115**, 104102 (2014).
- Lee, S. & Randall, C. A. Determination of electronic and ionic conductivity in mixed ionic conductors: HiTEC and in-situ impedance spectroscopy analysis of isoalvalent and aliovalent doped BaTiO<sub>3</sub>. *Solid State Ion.* **249–250**, 86–92 (2013).
- Morozovska, A. N. *et al.* Thermodynamics of electromechanically coupled mixed ionic-electronic conductors: deformation potential, Vegard strains, and flexoelectric effect. *Phys. Rev. B* **83**, 195313 (2011).
- Poumellec, B., Marucco, J. F. & Lagnel, F. Electron transport in Ti<sub>1–x</sub>Nb<sub>x</sub>O<sub>2</sub> solid solutions with  $x < 4\%$ . *J. Phys. Chem. Solids* **47**, 381–385 (1986).
- Narvaez, J., Saremi, S., Hong, J., Stengel, M. & Catalan, G. Large flexoelectric anisotropy in paraelectric barium titanate. *Phys. Rev. Lett.* **115**, 037601 (2015).
- Biancoli, A., Fancher, C. M., Jones, J. L. & Damjanovic, D. Breaking of macroscopic centric symmetry in paraelectric phases of ferroelectric materials and implications for flexoelectricity. *Nat. Mater.* **14**, 224–229 (2015).
- Garten, L. M. & Troler-McKinstry, S. Enhanced flexoelectricity through residual ferroelectricity in barium strontium titanate. *J. Appl. Phys.* **117**, 094102 (2015).
- Haertling, G. H. Rainbow actuators and sensors: a new smart technology. *Proc. SPIE* **3040**, 81–92 (1997).
- Berlincourt, D. & Jaffe, H. Elastic and piezoelectric coefficients of single-crystal barium titanate. *Phys. Rev.* **111**, 143–148 (1958).

**Acknowledgements** This research was funded by an ERC Starting grant from the EU (ERC 308023) and by a national research grant (FIS2013-48668-C2-1-P) from the Spanish MINECO. All research in ICN2 is supported by the Severo Ochoa Excellence Programme (SEV-2013-0295). F.V.-S. thanks MICITT and CONICIT for support during his PhD. We thank D. Torres for the illustration in Fig. 1, F. Belarre for help with sample polishing and B. Ballesteros for help with the EELS measurements shown in Methods.

**Author Contributions** J.N. and G.C. conceived the idea and analysed the results. J.N. and F.V.-S. performed the experiments. G.C. wrote the paper.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.C. ([gustau.catalan@icn2.cat](mailto:gustau.catalan@icn2.cat)).

**Reviewer Information** Nature thanks E. Eliseev, N. Mathur, D. Vanderbilt and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Materials.** The samples used in this work were as follows. (I) Single crystals of (001)-oriented BTO, commercially acquired from SurfaceNet, with dimensions of 1 mm × 10 mm × 0.46 mm (width × length × thickness). Reduction of the BTO<sub>-δ</sub> sample was achieved by annealing for 2 h at a temperature of 900 °C inside a vacuum chamber at a base pressure of 10<sup>-6</sup> mbar. The sample was heated at 10 °C min<sup>-1</sup> up to the annealing temperature; after annealing it was quenched back to room temperature in vacuum by turning off the heater, resulting in an exponential cooling rate with a time constant of 15 min. The re-oxidized sample (BTO<sub>+</sub><sub>δ</sub>) was made by re-annealing at 800 °C for 30 h in a pure (99.9999%) O<sub>2</sub> atmosphere. (II) Undoped and 0.05% Nb-doped (by weight) single crystals of rutile TiO<sub>2</sub> were acquired from Shinkosha (<http://www.shinkosha.com>) and polished by us to different thicknesses. The dimensions were 20 mm × 2 mm × 0.175–0.53 mm (length × width × thickness).

**Flexoelectric and dielectric measurements.** The method for measuring the flexoelectric coefficient is described in ref. 26: a dynamic mechanical analyser (Perkin-Elmer DMA-8000) was used to deliver an oscillatory three-point bending stress with 2 μm of vertical displacement to beam-shaped single-crystal samples placed between two supporting edges separated by 8 mm. The bending-induced displacement currents were measured using a lock-in amplifier (Stanford Research SR830 DSP) and converted to charge density by dividing the measured current by the frequency of the applied mechanical stress and by the area of the electrodes. The frequency of the oscillatory force in our dynamic mechanical analyser (DMA) was set to 13 Hz; a prime number was chosen to avoid ringing interference from the main power lines. The capacitance and dielectric loss were measured at 1 kHz with an Agilent LCR-meter model E4980A.

**Calculation of surface piezoelectricity.** By inverting Equation (1) we isolate the transverse surface piezoelectric coefficient:  $e_{13}^{\text{surf}} = 2\mu_{13}^{\text{eff}}/t$ , where  $t$  is the thickness of the crystal and  $\mu_{13}^{\text{eff}}$  is the experimentally measured effective flexoelectric coefficient of the semiconducting crystals (Fig. 3); please note that Equation (1) is only valid if bulk flexoelectricity is internally screened by the conductivity of the crystal.

Here  $e_{13}^{\text{surf}}$  is the strain–polarization piezoelectric coefficient of the surface, which can be converted into the more familiar force–charge piezoelectric coefficient ( $d_{ij}$ , with tensorial subindices  $i, j$ , indicating respectively the type of strain in Voigt notation and the direction of polarization in Cartesian coordinates) multiplying by the in-plane elastic compliance,  $s_{11}$  (ref. 31):

$$d_{13}^{\text{surf}} = s_{11} \frac{2\mu_{13}^{\text{eff}}}{t} \quad (2)$$

Using the results in Fig. 3 and the in-plane elastic compliance  $s_{11} \approx 8.5 \times 10^{-12} \text{ Pa}^{-1}$  for barium titanate<sup>30</sup>, we obtain the surface piezoelectric coefficient of BTO<sub>-δ</sub>, which ranges from a peak value of 37 pC N<sup>-1</sup> at  $T_C = 125^\circ\text{C}$  to 0.6 pC N<sup>-1</sup> at 300 °C. That this surface piezoelectricity is measured above the Curie temperature of BTO and so cannot be due to bulk piezoelectricity.

**Thickness dependence of the effective flexoelectric coefficient.** Although we did not manage to measure BTO crystals of different thicknesses (our thinned-down samples snapped when going through the ferroelectric–ferroelastic transition), we were able to measure Nb-doped TiO<sub>2</sub> of different thicknesses. The results are shown in Extended Data Fig. 1. The effective flexoelectric coefficient grows in proportion to the crystal thickness, which rules out a bulk mechanism (either flexoelectricity or inverse-Vegard electrochemical strain<sup>24</sup>). More data are required for accurate quantification of the linearity; for our limited sample set, the effective flexoelectric coefficient appears to grow at a rate faster than linear. Nevertheless, a linear regression to the available results (red line in Extended Data Fig. 1) yields a slope that can be introduced into Equation (2) to yield the surface piezoelectric coefficient. The result is  $d_{13}^{\text{surf}} = 16 \text{ pC N}^{-1}$ —a reasonable value that is between the piezoelectric coefficients of BTO and quartz.

**Thickness of the skin layer.** Using a series capacitor model, we estimate an upper bound for the skin layer thickness. In the barrier-layer capacitor model, the capacitance of the conductive sample corresponds to the interfacial barrier layers, with the conductive core of the crystal acting functionally as an intercalated electrode. Therefore, the capacitance of the conductive crystal is

$$C_{\text{BTO}-\delta} = \frac{1}{2} \varepsilon_i \frac{A}{t_i}$$

where  $A$  is the electrode area,  $t_i$  is the thickness of the interfacial barrier layer, and  $\varepsilon_i$  is its dielectric constant; the factor of 1/2 comes from assuming that there are two identical interfacial barrier layers with their capacitances added in series (that is, reciprocally).

The capacitance of the fully oxidized (dielectric) sample is given by  $1/C_{\text{BTO}} = 2/C_i + 1/C_b$ , where  $C_b$ ,  $C_i$  are the capacitances of the bulk and the two

interfacial layers. If we assume that the capacitance of the interface is much larger than that of the bulk—which is a reasonable assumption given that capacitance is inversely proportional to thickness and the interfacial layers are expected to be much thinner than the bulk—then we can simplify this expression by neglecting the interfacial contribution, so that:  $C_{\text{BTO}} = \varepsilon_b A/t_b$ , where  $\varepsilon_b$  and  $t_b$  are respectively the dielectric constant of the bulk and its thickness. Therefore, the ratio between the capacitance of the conducting crystal and that of the insulating crystal is proportional to the thickness ratios between the bulk and the interfacial barrier layers:

$$\frac{t_i}{t_b} = \frac{1}{2} \frac{\varepsilon_i}{\varepsilon_b} \frac{C_{\text{BTO}}}{C_{\text{BTO}-\delta}}$$

where  $t_{ib}$  and  $\varepsilon_{ib}$  are the thicknesses and relative dielectric constants, respectively, of the interface ('i') and the bulk ('b'). The crystal has a thickness of 460 μm, and the effective capacitance of the conducting crystal is approximately 200 times larger than that of the insulating crystal (see Fig. 2); therefore, the thickness of each interfacial barrier can be, at most,  $\lambda = 1/2 \times 460/200 = 1.15 \mu\text{m}$ . Figure 2 assumes that the dielectric constant of the barrier layer is similar to that of the bulk. If the local dielectric constant of the interfacial barrier layer was smaller than that of the bulk (which would be unusual in ferroelectrics, in which the interface behaves as a 'dead layer'), then the result must be corrected by a factor of  $\varepsilon_{\text{interface}}/\varepsilon_{\text{bulk}}$ . Hence,  $\lambda \approx 1 \mu\text{m}$  is an upper bound, with the real thickness of the barrier layer likely to be smaller.

**Re-oxidation kinetics and oxygen content.** Although sample preparation was done in vacuum, the flexoelectric measurements were done in a DMA with an air atmosphere and with temperatures reaching up to 300 °C; so, partial re-oxidation of the sample is possible.

The re-oxidation kinetics of ceramic BaTiO<sub>3</sub> doped with Ho<sub>2</sub>O<sub>3</sub> are described in ref. 32. The lowest measurement temperature therein is higher than our maximum measurement temperature: 450 °C versus 300 °C; however, we can extrapolate results of ref. 32 to obtain an expected diffusion coefficient of  $D = 1.6 \text{ cm}^2 \text{ s}^{-1}$  at 300 °C. The diffusion length is  $\lambda = 2\sqrt{D\tau}$ , where  $\tau$  is time; therefore, to re-oxidize 1 μm (maximum estimated thickness of our barriers), the required time would be only 15 s, which is well within the timescale of our measurements. According to these calculations, it is possible that the barrier layer is re-oxidized and that this is the reason for its insulating properties.

However, the diffusivity reported in ref. 32 is for doped ceramics, which have lower activation energies and higher ionic conductivities than do undoped single crystals. If, instead, we extrapolate the results for undoped BaTiO<sub>3</sub> reported in ref. 33, then we obtain a much lower diffusivity at 300 °C:  $D = 8 \times 10^{-14} \text{ cm}^2 \text{ s}^{-1}$ . Such a low diffusivity would require 32,000 s (almost 9 h) for the oxygen to diffuse to a depth of 1 μm. Therefore, although surface re-oxidation is possible, given the large range of diffusion coefficients in the literature, we cannot state the depth of the re-oxidized region with certainty or, consequently, discern whether the insulating barrier is caused by local re-oxidation or by electronic charge depletion (that is, Schottky barriers) at the semiconductor–electrode interface.

To shed more light on this question, we performed an Electron Energy Loss Spectroscopy (EELS) analysis of a cross-section of BTO<sub>-δ</sub> (Extended Data Fig. 2, top). The results do not show a monotonic trend in the Ti L absorption peaks that are typically used to characterize oxygen content in perovskite titanates: the peaks shift initially to higher energies (consistent with higher oxidation<sup>34,35</sup>), going from a depth of 1.4 μm to 150 nm, but then shift back to lower energies (consistent with lower oxidation) at a depth of 10 nm. A comparison with the L<sub>3</sub> peak in the EELS spectra of oxygen-deficient SrTiO<sub>3-δ</sub> (ref. 34) and of BaTiO<sub>3-δ</sub> (ref. 35) (Extended Data Fig. 2, bottom) shows that the L<sub>3</sub> peak split, which is inversely related to oxygen content, is consistent with a vacancy concentration of  $0.07 \leq \delta \leq 0.14$ , or between 2.3 mol% and 4.6 mol%.

We can quantify vacancies more accurately using gravimetry experiments. Using a high-precision analytical balance (Sartorius CPA225D), we measured the weight of a BaTiO<sub>3</sub> single crystal before and after vacuum annealing. The initial weight of the sample (10 mm × 1 mm × 2 mm; length × thickness × width) was 123.12 mg. The weight after vacuum annealing was 122.14 mg. Assuming that the weight reduction is due to oxygen loss, the decrease in weight (0.98 mg) represents an oxygen loss of 3.9 mol% ( $\delta = 0.12$ ) which is in the range estimated by the EELS analysis.

**Reproducibility.** Measuring flexoelectricity involves bending brittle crystals that tend to break and are not easy to replace; repeating measurements on different crystals is thus not an easy task. Nevertheless, for the reduced BTO<sub>-δ</sub> sample, we managed to measure several heating and cooling cycles and obtained similar results for all of them (Extended data Fig. 3). For the conductive Nb-doped TiO<sub>2</sub> crystal, we also measured two samples with thicknesses of ~0.5 mm, yielding results that differed by less than 20%. These two data points are included in Extended Data Fig. 1.

**Electronic versus ionic origin of the enhanced flexoelectricity.** Oxygen-deficient BTO is a mixed ionic–electronic conductor. Consequently, oxygen vacancies can



simultaneously act as ionic charge carriers and as electron donors that change the valence of  $\text{Ti}^{+4}$  to  $\text{Ti}^{+3}$ , thus introducing electrons in the conduction band. There is therefore a question about the origin of the enhanced electromechanical response: is it due to electronic screening of the polar discontinuity between the barrier layer and the bulk, or is it due to the strain gradient 'squeezing' the vacancies towards the convex side of the crystal, thus creating a defect concentration gradient with associated space-charge polarization? Such a Vegard-like 'flexoionic' mechanism would be the converse of the well-known electrochemical strain response of soft ionic conductors such as nafion or muscle tissue, the theory of which is described in refs 24, 36.

To address this question, we examine two issues: (I) the ratio between ionic and electronic conductivity in our measured range of temperature and frequency; and (II) whether a flexoelectric enhancement can be observed in a purely electronic semiconductor.

First, let us examine the issue of electronic versus ionic conductivity in oxygen-deficient BTO. Electrons are lighter and more mobile than ions, so they are expected to dominate conductivity in ceramics at the relatively low temperatures (by ionic standards) in which we have performed our work. For BTO, even at temperatures as high as 750 °C, ionic conductivity represents only about 20% of the total conductivity<sup>37</sup>; at 300 °C (our maximum temperature), the ionic contribution is expected to be exponentially lower.

We attempted to quantify the actual ionic share of the conductivity in our samples by extrapolating the ionic conductivity of  $\text{BTO}_{-\delta}$  measured previously<sup>23</sup> at higher temperatures. Using equations (12) and (13) from ref. 23, and the vacancy concentration of  $\delta = 0.12$  that we measured (using EELS and gravimetry analysis), we obtain an ionic conductivity of  $\sigma_{\text{ion}} = 4 \times 10^{-5} \Omega^{-1} \text{cm}^{-1}$  at 300 °C. This ionic conductivity would be higher than that of YSZ at the same temperature, a material that is already considered to be a good oxygen-vacancy conductor (it is the industry standard). We believe that this value is probably an imprecise overestimate because it is based on an extrapolation of measurements from the temperature range 950–1,050 °C down to  $T \leq 300$  °C; an extrapolation range that is much bigger than the initial fitting range is likely to result in an inaccurate estimate. Even leaving this important caveat aside, the ionic conductivity would still represent less than 5% of the total conductivity of  $\text{BTO}_{-\delta}$  at 300 °C. (Extended data Fig. 4).

So, although  $\text{BTO}_{-\delta}$  is indeed a mixed ionic–electronic conductor, in the temperature range we use, it behaves mostly as an electronic semiconductor with barrier layers caused by interfacial band-bending<sup>21</sup>.

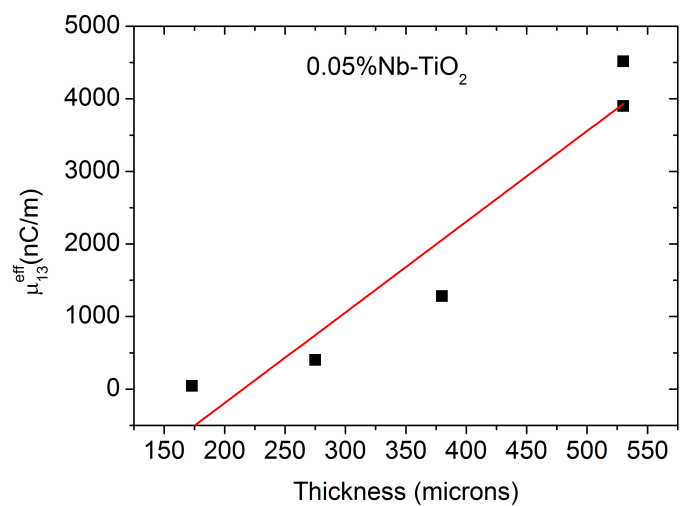
The ionic conduction model also seems to be inconsistent with the behaviour as a function of temperature. As temperature increases, ionic mobility increases exponentially, so a Vegard-like model would predict an exponential increase in bending-induced ionic drift—that is, increasing temperature should increase the bending-induced polarization. However,  $\text{BTO}_{-\delta}$  shows the opposite trend (Fig. 3 and Extended Data Fig. 3): above  $T_C$ , flexoelectricity decreases with temperature, showing a Curie–Weiss-like behaviour that is more consistent with a surface-piezoelectric model.

To determine whether enhanced flexoelectricity can be achieved in a purely electronic semiconductor, we performed measurements on Nb-doped  $\text{TiO}_2$  (0.05% Nb concentration by weight), which is a material for which the conductivity is purely electronic; each additional  $\text{Nb}^{+5}$  ion forces a  $\text{Ti}^{+4}$  ion to change its valence to  $\text{Ti}^{+3}$  to preserve charge neutrality, thus adding an electron to the conduction band<sup>25</sup>. The results are shown in Extended Data Fig. 5: Nb-doped  $\text{TiO}_2$  displays >2,000 times more effective flexoelectricity than the undoped (insulating)  $\text{TiO}_2$  crystal, suggesting that the effect is of electronic rather than ionic origin. This huge enhancement persists down to room temperature, at which vacancies, if there are any—and none are expected—should not be mobile enough to cross the 0.5-mm thickness of the crystal in less than 100 ms, as is required by our measurement frequency of 13 Hz.

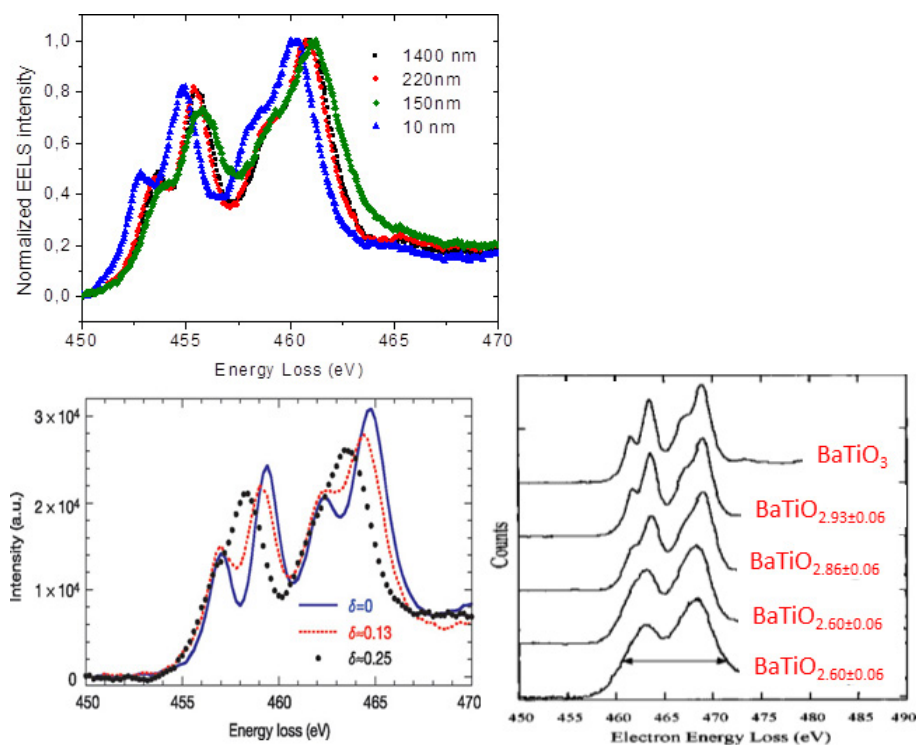
Therefore, our conclusion is that the enhancement of effective flexoelectricity for semiconductors does not require ionic transport.

**Inverse flexoelectricity.** The converse effect of bending induced by voltage can in principle also exist, but we believe that it will be much harder to observe in semiconductors than in insulators for at least two reasons. First, Joule heating caused by the leakage current across the bulk will cause the semiconductor to expand, and this thermal expansion can be larger than any other electromechanical response from the sample. Second, the most likely origin of the low-conductivity interfaces is charge depletion caused by band-bending (that is, Schottky barriers). In any rectifying junction, these barriers are wiped out by the application of an external voltage that is larger than the band misalignment between the electrode and the semiconductor, so the maximum voltage that one can apply before the barriers become conducting is limited by the semiconductor bandgap. In practice, this means that any external voltage that is larger than a couple of volts will essentially eliminate the barrier and turn it into an ohmic junction, thus removing the possibility of measuring its piezoelectric response.

31. Damjanovic, D. Ferroelectric, dielectric and piezoelectric properties of ferroelectric thin films and ceramics. *Rep. Prog. Phys.* **61**, 1267–1324 (1998).
32. Kaneda, K. *et al.* Kinetics of oxygen diffusion into multilayer ceramic capacitors during the re-oxidation process and its implications on dielectric properties. *J. Am. Ceram. Soc.* **94**, 3934–3940 (2011).
33. Müller, A. & Härdtl, K. H. Ambipolar diffusion phenomena in  $\text{BaTiO}_3$  and  $\text{SrTiO}_3$ . *Appl. Phys. A* **49**, 75–82 (1989).
34. Muller, D. A., Nakagawa, N., Ohtomo, A., Grazul, J. L. & Hwang, H. Y. Atomic-scale imaging of nanoengineered oxygen vacancy profiles in  $\text{SrTiO}_3$ . *Nature* **430**, 657–661 (2004).
35. Yang, G. Y., Dickey, E. C., Randall, C. A., Randall, M. S. & Mann, L. A. Modulated and ordered defect structures in electrically degraded Ni– $\text{BaTiO}_3$  multilayer ceramic capacitors. *J. Appl. Phys.* **94**, 5990–5996 (2003).
36. Lee, A. A., Colby, H. H. & Kornyshev, A. A. Statics and dynamics of electroactuation with single-charge-carrier ionomers. *J. Phys. Condens. Matter* **25**, 082203 (2013).
37. Chan, N.-H., Sharma, R. K. & Smyth, D. M. Nonstoichiometry in undoped  $\text{BaTiO}_3$ . *J. Am. Ceram. Soc.* **64**, 556–562 (1981).



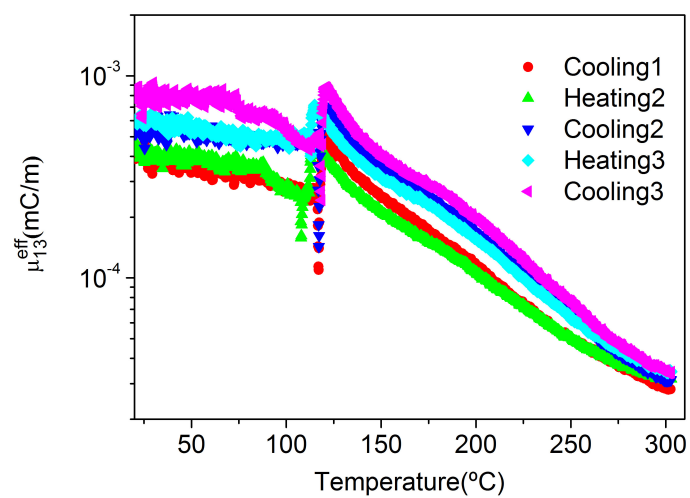
**Extended Data Figure 1 | Effective flexoelectric coefficients of semiconducting crystals of Nb-doped TiO<sub>2</sub> (0.05%Nb by weight) as a function of sample thickness. The red line is a linear fit to the data.**



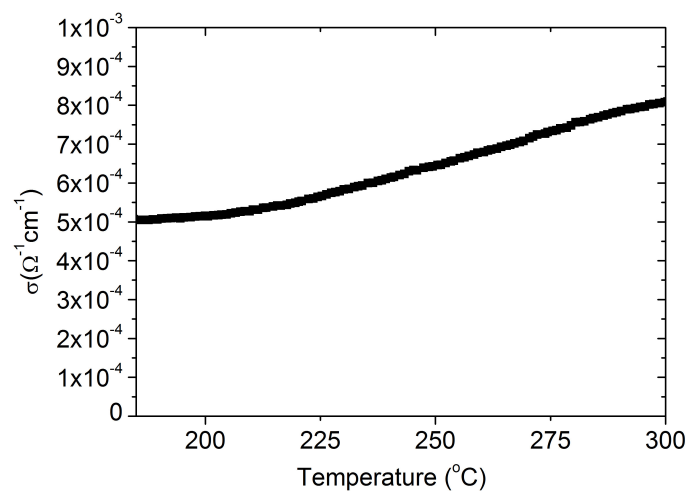
**Extended Data Figure 2 | EELS analysis.** Top, EELS spectra of a cross-sectional sample of BaTiO<sub>3</sub>, measured in a transmission electron microscope. There is no monotonic trend as a function of distance to the surface, so no indication that the surface (at least to a depth of 1.4  $\mu\text{m}$ ) is any more (or less) oxidized than the bulk. A comparison with the shape of

the EELS spectra of SrTiO<sub>3-δ</sub> (bottom-left; image reproduced from ref. 34, Macmillan Publishers Limited) or BaTiO<sub>3-δ</sub> (bottom-right; reprinted from ref. 35, with the permission of AIP Publishing) is consistent with  $\delta \leq 0.14$  for our crystals.

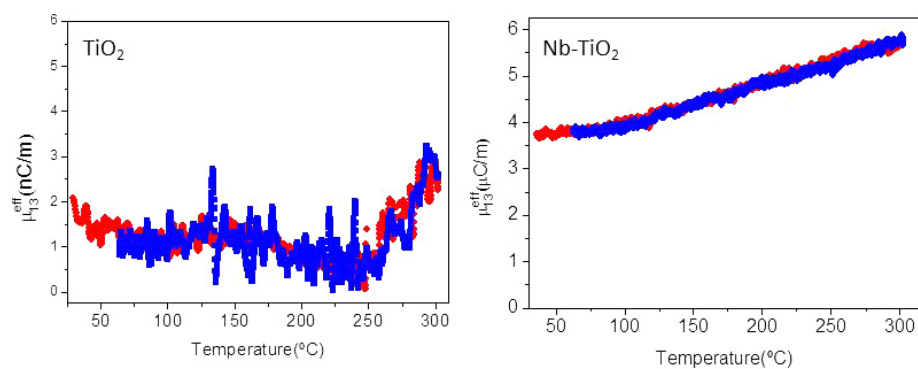




Extended Data Figure 3 | Consecutive measurements of the flexoelectric coefficient for semiconducting  $\text{BaTiO}_{3-\delta}$ .



**Extended Data Figure 4 | Conductivity of  $\text{BaTiO}_{3-\delta}$ .** Total conductivity  $\sigma = \sigma_{\text{electron}} + \sigma_{\text{ion}}$  measured across the capacitor structure.



**Extended Data Figure 5 | Flexoelectricity of undoped  $\text{TiO}_2$  and Nb-doped  $\text{TiO}_2$ .** The conducting Nb-doped sample (right) displays an effective flexoelectricity that is  $>2,000$  times larger than the insulating sample (left). Note that the units are  $\text{nC m}^{-1}$  and  $\mu\text{C m}^{-1}$  for  $\text{TiO}_2$  and Nb- $\text{TiO}_2$  respectively.



# Molecular transport through capillaries made with atomic-scale precision

B. Radha<sup>1</sup>, A. Esfandiar<sup>1</sup>, F. C. Wang<sup>2</sup>, A. P. Rooney<sup>3</sup>, K. Gopinadhan<sup>1</sup>, A. Keerthi<sup>1</sup>, A. Mishchenko<sup>1</sup>, A. Janardanan<sup>1</sup>, P. Blake<sup>4</sup>, L. Fumagalli<sup>1,4</sup>, M. Lozada-Hidalgo<sup>1</sup>, S. Garaj<sup>5</sup>, S. J. Haigh<sup>3</sup>, I. V. Grigorieva<sup>1</sup>, H. A. Wu<sup>2</sup> & A. K. Geim<sup>1</sup>

Nanometre-scale pores and capillaries have long been studied because of their importance in many natural phenomena and their use in numerous applications<sup>1</sup>. A more recent development is the ability to fabricate artificial capillaries with nanometre dimensions, which has enabled new research on molecular transport and led to the emergence of nanofluidics<sup>2–4</sup>. But surface roughness in particular makes it challenging to produce capillaries with precisely controlled dimensions at this spatial scale. Here we report the fabrication of narrow and smooth capillaries through van der Waals assembly<sup>5</sup>, with atomically flat sheets at the top and bottom separated by spacers made of two-dimensional crystals<sup>6</sup> with a precisely controlled number of layers. We use graphene and its multilayers as archetypal two-dimensional materials to demonstrate this technology, which produces structures that can be viewed as if individual atomic planes had been removed from a bulk crystal to leave behind flat voids of a height chosen with atomic-scale precision. Water transport through the channels, ranging in height from one to several dozen atomic planes, is characterized by unexpectedly fast flow (up to 1 metre per second) that we attribute to high capillary pressures (about 1,000 bar) and large slip lengths. For channels that accommodate only a few layers of water, the flow exhibits a marked enhancement that we associate with an increased structural order in nanoconfined water. Our work opens up an avenue to making capillaries and cavities with sizes tunable to ångström precision, and with permeation properties further controlled through a wide choice of atomically flat materials available for channel walls.

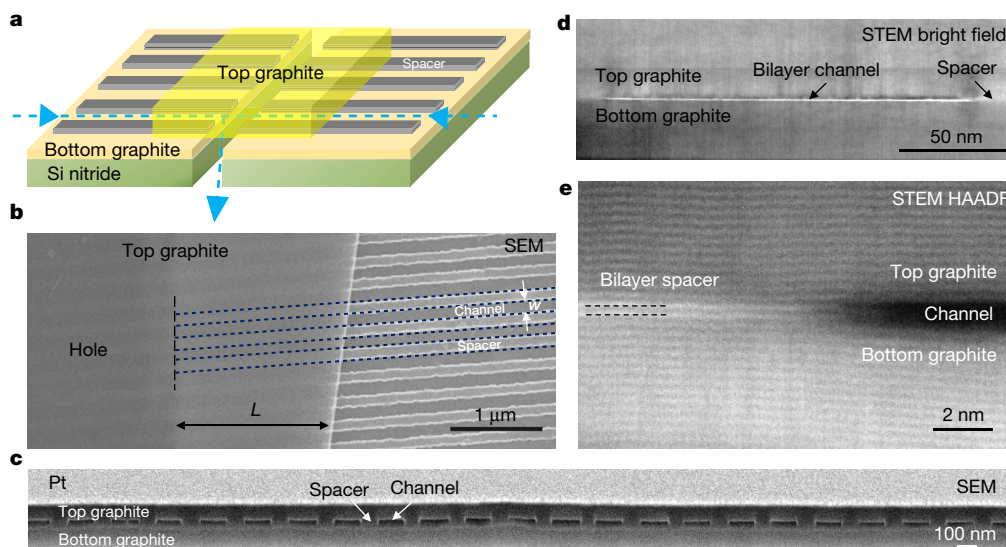
There are two principal routes for making pores and capillaries with nanometre dimensions<sup>7</sup>. The top-down approach uses micro- and nano-fabrication techniques and has realized channels down to 2 nm in average height<sup>8</sup>, but is fundamentally limited by surface roughness that is hard to reduce below a few nanometres using conventional materials and techniques<sup>9</sup>. The alternative bottom-up approach uses chemical synthesis with many advantages for scalable manufacturing, but which offers limited flexibility—especially for making capillaries with dimensions larger than several ångströms. Notable exceptions are nanotubes of carbon and other materials, which offered opportunities for studying mass transport through channels with nanometre diameters and atomically smooth walls<sup>10–17</sup> and promised new kinds of membrane and nanofluidic systems. But it has proved extremely difficult to integrate nanotubes into macroscopic devices, which perhaps explains the continuing controversy about fast water transport through carbon nanotubes (CNTs): conflicting findings from the very few experimental groups who succeeded in studying their permeation properties<sup>10–12,15,16</sup> have been discussed intensively in theoretical literature<sup>13,14,17</sup>, but with little further input from experiment. Graphene has also attracted considerable attention as a core material for making ultra-short nanopores<sup>18–23</sup>, and gas, liquid, ion and DNA transport through such pores has been reported. But the fundamental

restrictions inherent to top-down and bottom-up techniques also limit the ability to control the diameters of graphene nanopores precisely. We overcome such problems by exploiting both the atomic flatness of graphene (which allows for relatively long channels with atomically smooth walls, somewhat similar to CNTs) and its atomic thinness (which, through stacking, provides atomic-scale control of the channel's principal dimension, height). Our approach also preserves much of the flexibility offered by microfabrication techniques.

Figure 1a explains the basic idea behind our nanocapillary devices: they consist of atomically flat top and bottom graphite crystals that are separated by an array of spacers made from few-layer graphene. Such structures are fabricated by van der Waals (vdW) assembly using dry transfer techniques<sup>5</sup> and a free-standing Si nitride membrane with a rectangular hole as mechanical support for the assembly. Figure 1b–d shows micrographs of some of our devices. For details of their fabrication, we refer to Methods section 'Making nanocapillary devices' and Extended Data Figs 1 and 2. We denote our devices by the number  $N$  of graphene layers used as spacers. The height  $h$  of the cavity available for molecular transport can then be estimated as  $Na$ , where  $a \approx 3.4$  Å is the interlayer distance in graphite, that is, the effective thickness of one graphene layer. All the capillaries reported here had the same channel width  $w \approx 130$  nm, and 200 of them were incorporated within each device to increase molecular flow (Fig. 1). Their length  $L$  varied from  $<2$  µm to  $\sim 10$  µm. Despite the large aspect ratios  $w/h$ , we found no sagging of the graphite walls, which would cause capillary closure (Fig. 1d and Extended Data Fig. 3).

Under ambient conditions, all surfaces are covered with various adsorbates including water and hydrocarbons<sup>24</sup>, and it is not unreasonable to expect that nanocapillaries could be blocked by contamination introduced during fabrication or adsorbed from the air. Accordingly, we first checked whether our devices were open for gas and ion transport. Extended Data Fig. 4 shows that this was the case and that He permeated through the capillaries. We carried out such He tests for practically all the devices and found them normally open, except for monolayer capillaries ( $N=1$ ), which never exhibited any detectable permeation. Devices with larger  $N$  gradually deteriorated and, after several days of measurements, often became blocked. We attribute this to a build-up of hydrocarbon contamination that creeps along surfaces and is present even under oil-free vacuum conditions in our He tests. On the other hand, if immersed in water, the capillaries showed much greater resilience. All the tested devices were found open (except for  $N=1$ , again) and exhibited ionic conductance scalable with their dimensions (see Methods section 'Ionic conductance' and Extended Data Fig. 5). Capillaries kept in water did not get blocked for months and could be repeatedly measured. The fact that such artificial channels with a height down to the ångström scale allow studies of molecular transport under normal (not ultra-high) vacuum conditions is perhaps the most surprising finding of this work.

<sup>1</sup>School of Physics and Astronomy, University of Manchester, Manchester M13 9PL, UK. <sup>2</sup>Chinese Academy of Sciences Key Laboratory of Mechanical Behavior and Design of Materials, Department of Modern Mechanics, University of Science and Technology of China, Hefei, Anhui 230027, China. <sup>3</sup>School of Materials, University of Manchester, Manchester M13 9PL, UK. <sup>4</sup>National Graphene Institute, University of Manchester, Booth Street East, Manchester M13 9PL, UK. <sup>5</sup>Department of Physics, National University of Singapore, 117542 Singapore.



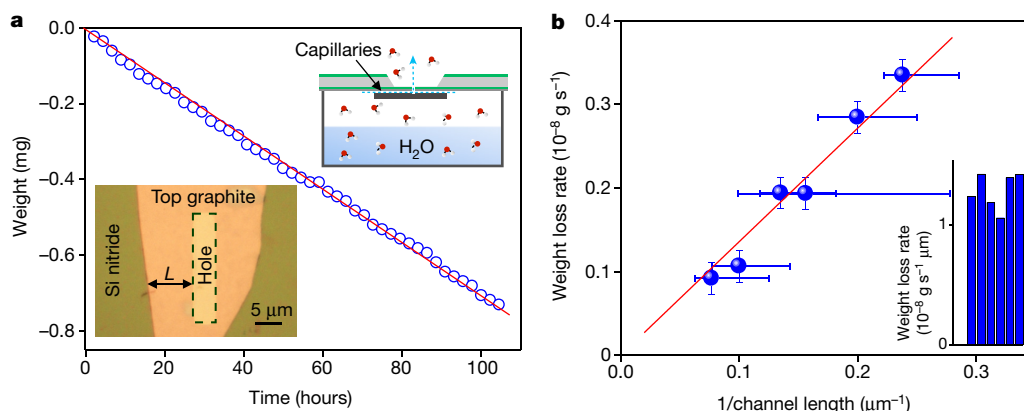
**Figure 1 | Graphene capillary devices.** **a**, General schematic of devices. The arrow indicates the flow direction used in all the experiments. **b**, Scanning electron microscopy (SEM) image of a trilayer device (top view). The spacers that are clearly seen in the area not covered by the top graphite can also be discerned underneath, running all the way to the hole etched in the bottom graphite. Three of the spacers are indicated by dotted lines and the edge of the hole by the dashed line. **c**, SEM

micrograph of a cross-section of another device showing an array of capillaries with cavity height  $h \approx 15$  nm. **d**, Cross-sectional bright field image of a bilayer capillary ( $h \approx 7$  Å) in a scanning transmission electron microscope (STEM). **e**, High-angle annular dark field (HAADF) image of the edge of the channel. The lamellae for cross-sectional imaging were made by focused ion beam milling (see Methods section ‘Visualization and characterization of graphene capillaries’).

Given the intense interest in nanoconfined water and the high stability of our devices in water, we explored their properties with respect to water permeation using precision gravimetry. As sketched in Fig. 2a and described in full in Methods, we measured weight loss from a miniature container that was filled with water and sealed with a Si nitride chip incorporating a nanocapillary device (Fig. 2a inset and Extended Data Fig. 6). An example of such measurements is shown in Fig. 2a. The slope of the measured curve yields the water evaporation rate,  $Q$ . Because the total cross-section of our devices is typically  $< 0.1 \mu\text{m}^2$ , measurements with microgram precision over several days were required to achieve accurate determinations of  $Q$ . Figure 2b shows  $Q$  observed for six devices with the same height ( $N=3$ ) but different  $L$ . Within our accuracy,  $Q$  was found to vary proportionally with  $1/\tilde{L} = \langle 1/L \rangle$ , where  $\tilde{L}$  is the effective average length with respect to a viscous flow, and  $\langle \dots \rangle$  denotes averaging over contributions from channels with different  $L$  (see equation (1)). This dependence on  $L$  unambiguously indicates that the observed evaporation rate was limited by water flow through capillaries, in agreement with additional

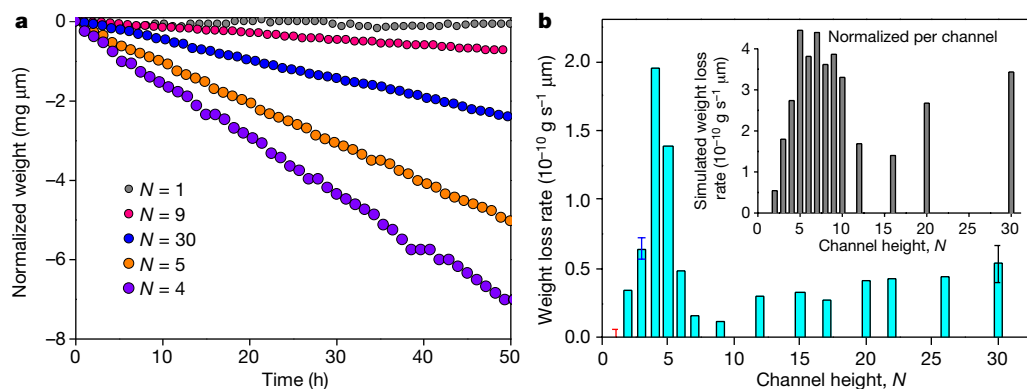
experimental observations described in Methods section ‘Gravimetric measurements’. The reproducibility of our gravimetry results can be judged from the scatter in the inset of Fig. 2b where  $Q$  values for the  $N=3$  devices are normalized by their  $\tilde{L}$ . All the trilayer capillaries show practically the same  $Q \approx 10^{-8} \text{ g s}^{-1}$  normalized for  $1 \mu\text{m}$  length, which translates into a flow velocity of  $\sim 0.1 \text{ m s}^{-1}$  for the shortest device in Fig. 2b ( $\tilde{L} \approx 4 \mu\text{m}$ ). As a control, we fabricated devices following exactly the same fabrication procedures but without graphene spacers ( $N=0$ ), in which case no weight loss could be detected. In addition, we tested our gravimetric set-up using micrometre apertures made in Si nitride membranes and found evaporation rates that agree well with those expected from theory (see Methods section ‘Gravimetric measurements’).

Having proved the accuracy and reproducibility of our measurements using trilayer devices, we investigated how the capillary flow depended on  $N$  using more than 30 different devices. Figure 3 shows that, as  $h$  decreases from  $\sim 10$  nm (maximum height in our gravimetry experiments),  $Q$  also decreases, as generally expected. However,



**Figure 2 | Water permeation through graphene nanocapillaries.** **a**, Weight loss due to water evaporation through one of our trilayer devices. Bottom inset, optical image (natural colour) of this particular device that has 200 parallel channels with  $L$  ranging from  $3.6 \mu\text{m}$  to  $10.1 \mu\text{m}$ . Top inset, a basic schematic of our gravimetric measurements. **b**, Water

evaporation rate ( $Q$ ) measured for six trilayer devices with different effective  $\tilde{L}$  (symbols; see text for definition of  $\tilde{L}$ ). The error bars indicate the range of  $L$  within each device. Inset, same data but normalized by  $\tilde{L}$ . The heights of the bars correspond to the measured  $Q$ .



**Figure 3 | Water flow through channels of different height.** **a**, Examples of gravimetric measurements for various  $N$  (the number of graphene layers used as spacers). They were carried out at 21 °C in near zero humidity, and the curves are normalized for the effective length of the devices,  $\tilde{L}$ . **b**, Dependence of  $Q$  on capillary height (the data are normalized by  $\tilde{L}$  and given per one channel). The blue error bar shows the s.d. for the data from

for  $h < 2$  nm,  $Q$  unexpectedly shoots up by more than an order of magnitude with respect to the trend exhibited by large- $N$  capillaries, and a strong peak appears at  $N=4-5$  (Fig. 3b). Devices with monolayer spacers exhibited no detectable weight loss, similar to the case of  $N=0$  and in agreement with our He and ion-conductance tests.

The entire evaporation process involves several steps, including transport of water vapour to capillary mouths inside the container, viscous flow through the graphene capillaries and subsequent diffusion and evaporation of transported water into air. To find out which steps affect the observed permeation, we carried out additional experiments. When the container was weighed upside down so that the liquid was in direct contact with the entries of the capillaries, exactly the same  $Q$  was recorded as in the upright position (see Methods section ‘Gravimetric measurements’). This is not surprising because, at 100% relative humidity (RH) inside the container and the contact angle of water on graphite<sup>25</sup>  $\phi \approx 55^\circ-85^\circ$ , the channels should be filled with the liquid owing to capillary condensation<sup>1,25</sup>. Atomic force microscopy and Raman spectroscopy further confirmed that the water inside the capillaries was in a liquid state (see Methods). We also note that the water meniscus cannot reside inside the nanocapillaries, given that the observed  $Q$  values require a water surface area of  $\sim 1 \mu\text{m}^2$  (as follows from the Hertz–Knudsen equation), that is, one to two orders of magnitude larger than the total cross-sectional area of the capillaries in our devices. Therefore, evaporation of the transported liquid must take place outside the mouths of the capillaries, and suggests an evaporating extended meniscus, which has been extensively studied in the literature for the case of macroscopic capillaries<sup>26</sup>. For our nanoscale openings, the extended meniscus is likely to involve an atomically thin layer of absorbed water that extends over micrometre distances being driven by high spreading pressures<sup>1</sup> (Extended Data Fig. 7a). To assess the role of this water film in our case, we measured  $Q$  for different RH outside the container. Surprisingly, no difference was found with increasing external RH up to values close to the onset of capillary condensation (see Methods section ‘Gravimetric measurements’). This shows that  $Q$  is not limited by diffusion and evaporation processes outside the container, and that the limiting process in our system is instead liquid flow through the graphene channels—in agreement with the finding that  $Q$  depends only on the capillary parameters  $\tilde{L}$  and  $N$ .

For long and wide rectangular channels with  $w/h \gg 1$ , liquid flow driven by pressure  $P$  is described by

$$Q = \rho \frac{h^3}{12\eta} \left( 1 + \frac{6\delta}{h} \right) \frac{P_w}{L} \quad (1)$$

where  $\rho$  is the water density,  $\eta$  its viscosity and  $\delta$  the slip length. All these characteristics of nanoconfined water may depend on  $h$ . To find

out whether equation (1) can explain the observed water transport behaviour, we performed molecular dynamics (MD) simulations using typical parameters for water–water and water–carbon interactions (Methods). Our analysis shows that  $\delta$  is large ( $\sim 60$  nm) but does not vary much with  $h$  (Extended Data Fig. 7), in agreement with previous MD results for flat graphene surfaces<sup>13,14,17</sup>. Also, changes in  $\rho$  are found to be relatively minor, reaching 4% for our smallest channels (Extended Data Fig. 8). The viscosity  $\eta$  increases by a factor of 2 for  $N < 5$ , which reflects the fact that water becomes more structured under nanoconfinement<sup>27–29</sup>. Using these parameters in equation (1), we find that  $Q$  detected for our smallest capillaries requires  $P$  of the order of 1,000 bar. This is consistent with supporting transport measurements using containers pressurized at  $\sim 1.5$  bar (close to the maximum pressure that our membranes could withstand), which revealed no difference in  $Q$  (see Methods section ‘Gravimetric measurements’).

It is difficult to perform MD simulations of the capillary pressure exerted by evaporating an extended meniscus because the low density of vapour necessitates a prohibitively large simulation volume. Therefore, we introduce the following simplification. The extended curvature of the meniscus is determined by two spatial scales, its height  $h$  and length outside the capillary mouth (Extended Data Fig. 7a). The length is expected to depend on RH but this was not the case experimentally, which allows us to approximate the extended meniscus using only  $h$ . With reference to Extended Data Figs 7a and 8a, both extended and internal menisci should have approximately the same height and involve the same interaction of water with graphite. Therefore, in both cases,  $P$  can be approximated<sup>1,30</sup> as  $P_0 + \Pi \approx 2\sigma\cos(\phi)/h + \Pi$ , where the first term describes the pressure due to a curved meniscus (with  $\sigma \approx 72$  mN m<sup>−1</sup> the surface tension of water). Even for our largest channels,  $P_0$  exceeds 10 bar. The second term  $\Pi$  refers to the so-called disjoining pressure<sup>1,30,31</sup> that describes the water–surface interaction, which can dominate at the nanoscale but rapidly decreases with increasing  $h$ . Our MD simulations (see Methods section ‘Capillary pressure’) show that, for large  $N > 10$ ,  $P$  roughly follows the classical dependence  $P_0$ , with  $\phi \approx 80^\circ$ ; but the disjoining pressure becomes dominant at smaller  $N$ , reaching above 1,000 bar (Extended Data Fig. 8b). Combining the simulated  $P$  with the other flow characteristics found in our MD analysis, equation (1) yields the  $Q(N)$  dependence shown in the inset of Fig. 3b. It qualitatively reproduces our experimental findings, including the peak at small  $N$  and even its absolute value. The physics behind the non-monotonic dependence  $Q(N)$  can be understood as follows. At large  $N$ , the classical contribution  $P_0 \propto 1/h$  dominates and equation (1) yields the linear dependence  $Q \propto h$ , in agreement with the trend observed in Fig. 3b for  $h > 3$  nm. Evaluating equation (1) numerically in the classical limit ( $P = P_0$ ), we find  $Q \approx 10^{-10}$  g s<sup>−1</sup> μm for  $h = 10$  nm,



in agreement with the values in Fig. 3b. The marked increase in  $Q$  for small  $N$  is due to the rapidly rising disjoining pressure, whereas the final fall in  $Q$  for smallest  $N$  occurs due to a combined effect of decreasing  $h$  and increasing  $\eta$ , which both reduce  $Q$ , overtaking the rise in  $II$  at small  $h$ . Note that, if it were not for the large enhancement factor  $6\delta/h$  due to the low friction of water against graphene walls, the simulated flow would be well below our detection limit.

The agreement between our model and the experiment is striking, especially if we consider the approximation used for calculating  $P$  and the unresolved experiment–theory dispute<sup>13,14,17</sup> concerning water permeation through CNTs. Further work is required to fully understand the mechanisms involved, and, in particular, to model extended menisci with nanoscale dimensions. Finally, the observed closure of monolayer capillaries ( $N=1$ ) seems to be not an accidental effect. Our MD analysis reveals that such narrow cavities are intrinsically unstable and collapse due to vdW attraction between opposite graphite walls (Extended Data Fig. 9).

Our fabrication approach allows capillary devices to be prepared in which the channel height can be controlled with true atomic precision by choosing spacers of different two-dimensional crystals (such as graphene, boron nitride, molybdenum disulphide) and their combinations. One can also alter the chemical and physical characteristics of these capillaries (for example, change their hydrophilicity) by using different atomically flat crystals for channel walls. Furthermore, the availability of highly insulating materials such as boron nitride and mica allows the design of nanofluidic systems in which ionic or mass transport can be controlled by gate voltage. Our current devices transfer minute amounts of liquid, typical for nanofluidics, but it is feasible to increase the flow by many orders of magnitude using dense arrays of short (submicrometre) capillaries covering millimetre-sized areas, which could be of interest for nanofiltration, for example.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 15 February; accepted 9 August 2016.**

**Published online 7 September 2016.**

- Israelachvili, J. N. *Intermolecular and Surface Forces* 3rd edn (Academic, 2011).
- Eijkel, J. T. & van den Berg, A. Nanofluidics: what is it and what can we expect from it? *Microfluid. Nanofluidics* **1**, 249–267 (2005).
- Schoch, R. B., Han, J. & Renaud, P. Transport phenomena in nanofluidics. *Rev. Mod. Phys.* **80**, 839–883 (2008).
- Howorka, S. & Siwy, Z. Nanopore analytics: sensing of single molecules. *Chem. Soc. Rev.* **38**, 2360–2384 (2009).
- Geim, A. K. & Grigorieva, I. V. Van der Waals heterostructures. *Nature* **499**, 419–425 (2013).
- Novoselov, K. S. *et al.* Two-dimensional atomic crystals. *Proc. Natl Acad. Sci. USA* **102**, 10451–10453 (2005).
- Mijatovic, D., Eijkel, J. C. T. & van den Berg, A. Technologies for nanofluidic systems: top-down vs. bottom-up — a review. *Lab Chip* **5**, 492–500 (2005).
- Duan, C. & Majumdar, A. Anomalous ion transport in 2-nm hydrophilic nanochannels. *Nat. Nanotechnol.* **5**, 848–852 (2010).
- Duan, C., Wang, W. & Xie, Q. Review article: fabrication of nanofluidic devices. *Biomicrofluidics* **7**, 026501 (2013).
- Hinds, B. J. *et al.* Aligned multiwalled carbon nanotube membranes. *Science* **303**, 62–65 (2004).
- Majumdar, M., Chopra, N., Andrews, R. & Hinds, B. J. Nanoscale hydrodynamics: enhanced flow in carbon nanotubes. *Nature* **438**, 44 (2005).

- Holt, J. K. *et al.* Fast mass transport through sub-2-nanometer carbon nanotubes. *Science* **312**, 1034–1037 (2006).
- Thomas, J. A. & McGaughey, A. J. H. Reassessing fast water transport through carbon nanotubes. *Nano Lett.* **8**, 2788–2793 (2008).
- Falk, K., Sedlmeier, F., Joly, L., Netz, R. R. & Bocquet, L. Molecular origin of fast water transport in carbon nanotube membranes: superlubricity versus curvature dependent friction. *Nano Lett.* **10**, 4067–4073 (2010).
- Majumdar, M., Chopra, N. & Hinds, B. J. Mass transport through carbon nanotube membranes in three different regimes: ionic diffusion and gas and liquid flow. *ACS Nano* **5**, 3867–3877 (2011).
- Qin, X., Yuan, Q., Zhao, Y., Xie, S. & Liu, Z. Measurement of the rate of water translocation through carbon nanotubes. *Nano Lett.* **11**, 2173–2177 (2011).
- Kannam, S. K., Todd, B. D., Hansen, J. S. & Dai, P. J. How fast does water flow in carbon nanotubes? *J. Chem. Phys.* **138**, 094701 (2013).
- Garaj, S. *et al.* Graphene as a subnanometre trans-electrode membrane. *Nature* **467**, 190–193 (2010).
- Koenig, S. P., Wang, L., Pellegrino, J. & Bunch, J. S. Selective molecular sieving through porous graphene. *Nat. Nanotechnol.* **7**, 728–732 (2012).
- Celebi, K. *et al.* Ultimate permeation across atomically thin porous graphene. *Science* **344**, 289–292 (2014).
- O'Hern, S. C. *et al.* Nanofiltration across defect-sealed nanoporous monolayer graphene. *Nano Lett.* **15**, 3254–3260 (2015).
- Wang, L. *et al.* Molecular valves for controlling gas phase transport made from discrete ångström-sized pores in graphene. *Nat. Nanotechnol.* **10**, 785–790 (2015).
- Jain, T. *et al.* Heterogeneous sub-continuum ionic transport in statistically isolated graphene nanopores. *Nat. Nanotechnol.* **10**, 1053–1057 (2015).
- Haigh, S. J. *et al.* Cross-sectional imaging of individual layers and buried interfaces of graphene-based heterostructures and superlattices. *Nat. Mater.* **11**, 764–767 (2012).
- Mücksch, C., Röscher, C., Müller-Renno, C., Ziegler, C. & Urbassek, H. M. Consequences of hydrocarbon contamination for wettability and protein adsorption on graphite surfaces. *J. Phys. Chem. C* **119**, 12496–12501 (2015).
- DasGupta, S., Schonberg, J. A., Kim, I. Y. & Wayner, P. C. Use of the augmented Young-Laplace equation to model equilibrium and evaporating extended menisci. *J. Colloid Interface Sci.* **157**, 332–342 (1993).
- Hummer, G., Rasaiah, J. C. & Noworyta, J. P. Water conduction through the hydrophobic channel of a carbon nanotube. *Nature* **414**, 188–190 (2001).
- Raviv, U., Laurant, P. & Klein, J. Fluidity of water confined to subnanometre films. *Nature* **413**, 51–54 (2001).
- Zhao, W.-H. *et al.* Highly confined water: two-dimensional ice, amorphous ice, and clathrate hydrates. *Acc. Chem. Res.* **47**, 2505–2513 (2014).
- Mate, C. M. Taking a fresh look at disjoining pressure of lubricants at slider-disk interfaces. *IEEE Trans. Magn.* **47**, 124–130 (2011).
- Gravelle, S., Ybert, C., Bocquet, L. & Joly, L. Anomalous capillary filling and wettability reversal in nanochannels. *Phys. Rev. E* **93**, 033123 (2016).

**Acknowledgements** This work was supported by the Lloyd's Register Foundation, the European Research Council and the Royal Society. B.R. and K.G. acknowledge Marie Curie International Incoming Fellowships. S.J.H. and A.P.R. acknowledge the EPSRC NowNANO programme, EP/M022498/1, EP/K016946/1 and DTRA (HDTRA1-12-1-0013) for funding.

**Author Contributions** A.K.G. and B.R. designed and directed the project. B.R. led development of the fabrication techniques. H.A.W. and F.C.W. provided theory support. B.R., A.E., K.G., A.J. and A.K. performed measurements and analysed results. A.P.R., S.J.H. and L.F. provided electron microscopy imaging. A.K.G., B.R., I.V.G., H.A.W. and F.C.W. wrote the manuscript. All authors contributed to discussions.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.R. (radha.boya@manchester.ac.uk), F.C.W. (wangfc@ustc.edu.cn) or A.K.G. (andre.k.geim@manchester.ac.uk).

**Reviewer Information** *Nature* thanks L. Bocquet, C. Duan, J. Eijkel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Making nanocapillary devices.** Our fabrication procedures are explained in Extended Data Fig. 1. First, we prepare a free-standing Si nitride membrane of approximately  $100 \times 100 \mu\text{m}^2$  in size using commercially available Si wafers with 500 nm thick Si nitride<sup>32</sup>. A rectangular hole ( $3 \times 20 \mu\text{m}^2$ ) is made in the membrane using the standard photolithography and reactive ion etching (step 1). Then a relatively thick ( $>10$  nm) graphite crystal is deposited to seal the opening (step 2) using the dry transfer method described in Supplementary Information of ref. 33. On a separate Si wafer (with 300 nm of  $\text{SiO}_2$ ) we prepare multilayer graphene of a chosen thickness using micromechanical cleavage<sup>6</sup> to serve as a spacer. The graphene crystal is patterned by electron beam lithography and oxygen plasma etching to create an array of parallel stripes of  $\sim 130$  nm in width and separated by the same distance (Fig. 1a). These dimensions are chosen to obtain sufficiently narrow channels (to prevent them from collapsing; see below) and, at the same time, to ensure full reproducibility using our lithography facilities. The graphene stripes are then transferred onto the bottom graphite so that they are aligned perpendicular to the long side of the rectangular opening (step 3). Oxygen plasma etching is employed to drill through the graphite–graphene stack using the hole in Si nitride as a mask (step 4). Finally, another graphite crystal (approximately 100 nm in thickness) is ‘dry-transferred’ to serve as the capping layer. This completes a set of graphene capillaries, such that their entries and exits are accessible from the opposite sides of the Si wafer (step 5). After each transfer, the assembly is annealed at  $400^\circ\text{C}$  for 3 h to remove possible contamination.

**Visualization and characterization of graphene capillaries.** In addition to Fig. 1, Extended Data Fig. 2 provides further examples of imaging of our graphene capillaries including their optical, atomic force microscopy (AFM), SEM and STEM micrographs. We used SEM and optical images such as in Fig. 1b and Extended Data Fig. 2a to measure lengths of our devices and calculate their average length with respect to frictional flow,  $\tilde{L} = (1/L)^{-1}$ . For most of our devices,  $L$  varied by less than 30% (for example, Extended Data Fig. 2a) and, accordingly, we found no qualitative difference if using  $\tilde{L}$  or  $\langle L \rangle$  in our analyses.

To obtain the cross-sectional SEM images shown in the figure, we used a dual-beam system (Zeiss Crossbeam 540), which combines electron microscopy with focused ion beam (FIB) capabilities. The region of interest was located using SEM, and a protective Pt layer ( $\sim 0.5 \mu\text{m}$  thick) was deposited on top. Then a trench was milled using 30 kV  $\text{Ga}^+$  beam at 0.1 nA current, which exposed the device’s cross-section. Two additional polishing steps at 10 and 1 pA were subsequently carried out using the same 30 kV  $\text{Ga}^+$  beam. During the final step, the raster had a width of 200 nm, and it took approximately 30 min to complete the polishing.

Samples for STEM were obtained by implementing the *in situ* lift-out procedures<sup>24,33,34</sup> in a FIB system (Helios Nanolab DualBeam 660). A cross-sectional lamella (that is, a thin foil cut out perpendicular to the capillary axes) was prepared by FIB milling and lifted from the substrate using a micromanipulator, aided by ion beam deposition of platinum. After transfer to a specialist OmniProbe grid, the foil was thinned down to  $<100$  nm and then polished to electron transparency using 5 kV and subsequently 2 kV ion milling. High-resolution STEM images were acquired in an aberration-corrected microscope (FEI Titan G2 80–200 kV) using a probe convergence angle of 21 mrad, a HAADF inner angle of 48 mrad and a probe current of  $\sim 80$  pA. To ensure that the electron probe was parallel to graphite planes, the cross-sectional sample was aligned to the relevant Kikuchi bands of the Si substrate and graphite.

To prevent closure of our nanocapillaries through sagging of their walls, it is essential to choose appropriate values for the channel width,  $w$ , and top graphite’s thickness,  $H$  (bottom graphite is supported by the substrate which stops it from sagging). To illustrate the crucial role of  $H$ , Extended Data Fig. 3 shows AFM images of trilayer channels covered with top graphite of a varying thickness. One can see that thin graphite ( $H \approx 12$  nm) sags—at least partially—into the channels whereas the thicker layer (52 nm) remains atomically flat, which suggests that the channels underneath are likely to remain open. For our standard channels with  $w \approx 130$  nm, we find that it requires  $H > 50$  nm to avoid their collapse. On the other hand, for  $w > 500$  nm, the top graphite crystal in our experiments always sagged into the channels (even for  $H > 200$  nm; Extended Data Fig. 3b).

To confirm the presence of liquid water inside our nanocapillaries, we carried out their Raman spectroscopy. A working device with  $N = 30$  was fabricated following the procedures described above but, instead of graphite, hexagonal boron nitride was used as the top layer. This provided optical access for a laser beam whereas hexagonal boron nitride’s mechanical and tribological properties are similar to those of graphite. We chose the large  $N$  to increase the amount of examined water as its Raman signal is relatively weak. If liquid water was placed inside the container, a clear signature of water (peak at  $\sim 3,250 \text{ cm}^{-1}$  using a laser wavelength of 785 nm) appeared in the Raman spectra taken from the area with graphene capillaries. No signal could be detected at low RH or outside the area. It is also instructive to mention that capillaries with small sagging ( $\leq 0.5$  nm)

reacted to high RH in such a way that the sagging disappeared and the top graphite layer became flat in the AFM images. For example, for capillaries with  $N = 5$  this straightening of graphene walls happened at  $\sim 70\%$  RH, indicating the onset of capillary condensation<sup>1</sup>. This allows an estimate for the contact angle  $\phi \approx 55^\circ$ , in agreement with  $\phi$  observed for water on clean graphite surfaces<sup>25</sup>. No changes with increasing RH were observed for sufficiently thick top layers that exhibited no initial sagging and were used in the studied devices.

**Helium permeation.** To ensure that the fabricated capillaries are not blocked by sagging or contamination, we checked gas permeation through them using a helium-leak detector (INFICON UL200). A principal schematic of our experimental set-up is shown in Extended Data Fig. 4a. In short, a Si wafer with a capillary device is clamped between O-rings and separates two oil-free vacuum chambers. One of them is equipped with pressure gauges and a pump to allow control of the applied helium pressure  $P_a$  at the capillary entry. The other chamber is connected to the leak detector. We have found our graphene–Si nitride membranes sufficiently robust to withstand  $P_a$  up to 2 bar. Examples of our tests are shown in Extended Data Fig. 4b. Except for devices with  $N = 0$  and 1, all other nanocapillaries allowed He permeation.

Although discussion of gas transport through graphene nanocapillaries is beyond the scope of the present report, it is instructive to compare the observed He rates  $Q$  with those expected theoretically. For a channel with  $h$  much smaller than the mean free path  $l \approx 140$  nm for He atoms at the atmospheric pressure, their mass transport is described by the Knudsen formula<sup>35</sup>

$$Q = \alpha P_a \left( \frac{M_{\text{He}}}{2\pi RT} \right)^{1/2} wh \quad (2)$$

where  $M_{\text{He}}$  is the atomic mass of He. For narrow-slit channels, the transmission coefficient  $\alpha$  can be approximated<sup>35</sup> by  $\alpha \approx 5(h/L)$ . In the case of  $h = 15$  nm and  $P_a = 100$  mbar, equation (2) yields  $Q \approx 7 \times 10^{-13} \text{ g s}^{-1} \mu\text{m}$ , in good agreement with our measurements shown in Extended Data Fig. 4b for  $N \approx 45$ . On the other hand, smaller capillaries ( $N \leq 5$ ) are found to exhibit leak rates that are nearly two orders of magnitude higher than the rates expected from equation (2). Moreover, their  $Q$  were even greater than those found for 10 times higher channels (Extended Data Fig. 4b), contrary to general expectations. A similar enhancement of He flow was previously reported for sub-2-nm CNTs and attributed to the atomic smoothness of graphene walls<sup>12</sup>.

For the case of a water vapour driven by the difference in RH (differential pressure of 23 mbar), equation (2) yields evaporation rates of  $\sim 5 \times 10^{-15} \text{ g s}^{-1} \mu\text{m}$  and  $400 \times 10^{-15} \text{ g s}^{-1} \mu\text{m}$  for channels with  $N = 5$  and 45, respectively, which is 3–5 orders of magnitude smaller than the experimental values in Fig. 3b. Even if the vapour permeation is enhanced by two orders of magnitude, as observed for the He transport through capillaries with  $N = 5$ , this still leaves three orders of magnitude unaccounted for. This disagreement provides yet another indication that water permeates through our graphene channels as a liquid.

**Ionic conductance.** We also tested a number of capillary devices using the electrochemical set-up shown in Extended Data Fig. 5a. KCl solutions of different concentrations  $C$  were introduced into two reservoirs separated by a Si wafer incorporating a graphene device under investigation. Possible air bubbles were removed by extensive flushing from both sides of the Si wafer. Current–voltage ( $I$ – $V$ ) characteristics were recorded using Keithley 2636 A SourceMeter and Ag/AgCl electrodes. Extended Data Fig. 5b, c shows examples of our measurements for two devices with  $N = 2$  and 17. The  $I$ – $V$  curves are linear at low biases and exhibit little hysteresis. At high  $C$ , the observed ionic currents for a given voltage differ approximately by a factor of  $\sim 8$ , in good agreement with the ratio between the channel heights. Devices with  $N = 0$  and 1 exhibited no detectable ionic conductance.

Extended Data Fig. 5d shows that the ionic conductance,  $G$ , increases linearly with  $C$  for ionic concentrations higher than  $10^{-2}$  M, and its absolute value agrees well with the values expected from the known bulk conductivity of KCl solutions. In the low concentration regime ( $<10^{-3}$  M),  $G$  saturates to a constant value, the same for both devices. Such saturation is typical for nanocapillaries and attributed to the surface charge effect<sup>8,36</sup>. In our case, the saturation value is very small and, taking into account electro-osmotic and finite- $\delta$  contributions<sup>37</sup>, we find a surface charge density of  $\sim 3 \times 10^{10} \text{ cm}^{-2}$ , orders of magnitude lower than the values observed for other nanocapillaries including CNTs<sup>38</sup>. This serves as another indication that graphene walls of our channels are impurity-free, in agreement with low charge densities usually found in graphene-based vdW heterostructures<sup>5</sup>.

**Gravimetric measurements.** The set-up used in our studies of water permeation is shown in Extended Data Fig. 6a, b. The assembled capillary device was mounted on top of a container partially filled with deionized water. The container was then placed on a microbalance (Mettler Toledo XPE26) and weighed in an enclosure with a constant temperature (typically,  $21 \pm 0.1^\circ\text{C}$ ) and at near 0% humidity that

was maintained using molecular sieves. The weight of the container was recorded at regular intervals (typically, 1 min) using a computer.

To verify the accurate operation of the gravimetric set-up, we prepared reference devices with round apertures of different diameters,  $D$ , etched in Si nitride membranes. Using the same sample mounting and measurements procedures as for our graphene devices, we measured water evaporation through the apertures (Extended Data Fig. 6). The Knudsen numbers for our apertures are small and the evaporation can be described by diffusion of water molecules through air. The molecular flow  $F$  is given by<sup>35</sup>

$$F = \frac{1}{3} \langle v \rangle l \frac{dn}{dx}$$

where  $\langle v \rangle$  is the average velocity of molecules in air,  $l \approx 60$  nm is the mean free path, and  $dn/dx$  the concentration gradient. To leave the container, water molecules have to diffuse through air over a distance of about  $D$ , which allows an estimate  $dn/dx \approx \Delta n/D$  where  $\Delta n = \Delta P/(k_B T)$  is the difference in water concentrations at large distances from the aperture and  $\Delta P$  the difference in their partial pressures. The diffusion problem can be solved exactly for the case of infinitely thin orifices, which is a reasonable approximation for our 500-nm-thick Si nitride membranes and yields<sup>39</sup>

$$\frac{dn}{dx} = \frac{4}{\pi} \frac{\Delta n}{D}$$

The resulting weight loss is given by

$$Q = FM_{H_2O} \frac{\pi D^2}{4} = \langle v \rangle l \frac{M_{H_2O}}{3k_B T} D \Delta P$$

where  $M_{H_2O}$  is the molecular weight of water. This equation yields  $Q \propto D$ , in agreement with the observed behaviour in Extended Data Fig. 6d. The counterintuitive linear dependence arises because the available area for diffusion increases proportionally to  $D^2$  whereas the diffusion length decreases as  $1/D$ . Using  $\Delta P = 23$  mbar, the above equation yields  $\sim 1.7 \times 10^{-10}$  g s<sup>-1</sup>  $\times D$  (in  $\mu$ m), which is within 15% from the best fit in Extended Data Fig. 6d. Importantly, the measurements for our aperture devices cover approximately the same range of  $Q$  as that found for graphene capillaries (Fig. 3). The excellent agreement between the experiment and theory confirms reliability of our gravimetry set-up.

To narrow down the range of possible explanations for the observed fast water flow, two additional sets of experiments were carried out. First, using devices with  $N = 5$  and 30, we increased RH outside the container up to 50% and >70%, respectively, using increments of 20%. Surprisingly, no changes in  $Q$  could be detected. At even higher external RH, the evaporation completely stopped (at  $\sim 90\%$  for  $N = 30$ ), which is attributed to condensation at the output side of the channels. This shows that it was not necessary to maintain RH accurately at zero and confirms once again that it was not the differential vapour pressure that drove the water flow. Most importantly, these experiments indicate that water diffusion and evaporation outside the capillaries was not a limiting factor in our permeation measurements. Otherwise, the increase in external RH would significantly reduce  $Q$ .

In the second set of experiments, we applied an additional pressure of  $1.3 \pm 0.3$  bar to the water column inside our containers. This pressure was chosen to be close to the maximum pressure that our membranes could withstand. To create such pressures while keeping the container weight below  $\sim 15$  g (required for precision gravimetry), a chosen amount of NaBH<sub>4</sub> was dissolved in water inside the container which resulted in a slow release of hydrogen (over several hours at room temperature). The pressure build-up inside a closed container was monitored in a separate experiment (without a graphene device) and quantitatively agreed with the pressure expected from the chemical reaction. The extra pressure did not lead to any discernible difference in  $Q$ . This unambiguously proves that  $P$  much higher than 1 bar drove water through the capillaries, and our measurement accuracy of  $\sim 10\%$  yields a lower bound estimate for  $P$  as 15 bar.

**Molecular dynamics simulations.** To understand the observed behaviour, we used both non-equilibrium and equilibrium MD simulations (NEMD and EMD, respectively). Water molecules were confined between two rigid graphene sheets of approximately  $5 \times 5$  nm<sup>2</sup> in size and separated by  $h = aN$  (Extended Data Fig. 7b). Unless specifically mentioned below, we used the SPC/E model for water<sup>40</sup>, and the carbon atoms were modelled as fixed neutral particles interacting with oxygen through the Lennard–Jones (LJ) potential with the standard values<sup>16,41,42</sup> of the interaction parameters,  $\epsilon_{CO}$  and  $\sigma_{CO}$ . For consistency, in all the presented simulations we used  $\epsilon_{CO} = 0.0927$  kcal/mol and  $\sigma_{CO} = 3.283$  Å, and LJ interactions were truncated using a cut-off of 10 Å. The temperature of water was maintained at 300 K using the Berendsen thermostat. Long-range Coulomb forces were computed using the particle-particle particle-mesh method, and all the simulations were carried out in the canonical ensemble using LAMMPS<sup>43</sup>. The graphene capillary shown

in Extended Data Fig. 7b was initially connected to two reservoirs that contained 5,000 water molecules each. A pressure of 1 bar was applied to the water reservoirs to ensure equal pressure on water molecules inside capillaries of different  $h$ . Then the reservoirs were removed and periodic boundary conditions were applied in all three directions.

In our NEMD analysis, the flow was generated by applying a constant unidirectional acceleration of  $10^{12}$  m s<sup>-2</sup> to all atoms in water, which corresponds to a pressure gradient of  $\sim 10^{15}$  Pa m<sup>-1</sup>. Such large gradients are standard for NEMD simulations and necessary to obtain statistically significant results<sup>44</sup>. The steady flow state was achieved after  $\sim 1$  ns, and the data were collected for  $> 10$  ns to find the streaming velocity  $V$ . The flux was calculated as  $Q = \rho w h V$  where  $\rho$  is the average density of the nanoconfined water and  $w$  the capillary width perpendicular to the flow direction. For the known  $Q$ , the slip length  $\delta$  can be found using equation (1). Our NEMD results are presented in Extended Data Fig. 7c.

We also used EMD simulations to find  $\delta = \eta/\lambda$  which is given by the ratio of the shear viscosity  $\eta$  to the liquid–solid friction coefficient,  $\lambda$ . Both  $\eta$  and  $\lambda$  were calculated through the Green–Kubo formalism using the simulated local structure of confined water<sup>45,46</sup>. We found that  $\eta$  was in the range of  $(0.5–0.9) \times 10^{-3}$  Pa s and  $\lambda$  was about  $10^4$  kg m<sup>-2</sup> s<sup>-1</sup>, in agreement with the previous simulations for the water–graphite interface<sup>14,44</sup>. This yielded  $\delta \approx 53 \pm 8$  nm for  $N$  ranging from 2 to 30 (Extended Data Fig. 7c). Note that CNTs are known to exhibit a strong dependence of  $\delta$  on their diameter<sup>14</sup>, which is attributed to the effect of curvature. No  $h$  dependence was found for planar graphene channels either in our simulations or previously<sup>14,17</sup>. For example, Falk *et al.*<sup>14</sup> reported  $\delta \approx 80$  nm for  $h$  ranging from 0.4 nm to 4 nm, and Kannam *et al.*<sup>17</sup> found  $\delta \approx 60 \pm 6$  nm for  $h \approx 4$  nm. The relatively minor discrepancies can be attributed to details of MD simulations such as different interaction parameters and different thermostats.

Despite usual<sup>44,45</sup> quantitative differences between NEMD and EMD simulations (Extended Data Fig. 7c), both show qualitatively the same behaviour with a rapid decrease in water flow with decreasing  $N$  (approximately,  $\propto h^2$  as expected from equation (1) for a constant pressure  $P$ ) and without any anomalies at small  $N$ . This is in agreement with the previously reported simulations for flat graphene capillaries<sup>17,44</sup>. We also tried other models for water (TIP4P/2005)<sup>47</sup> and its interaction with graphene<sup>48</sup> as well as the use of a flexible graphene confinement. However, if a pressure  $P$  was assumed independent of  $h$ , we found it impossible to obtain a peak in permeation at small  $N$ .

**Capillary pressure.** To analyse changes in the capillary pressure  $P$  with decreasing  $N$ , we used the MD set-up shown in Extended Data Fig. 8a. As discussed in the main text, the internal meniscus was chosen as an approximation for the evaporating extended meniscus sketched in Extended Data Fig. 7a. Both have approximately the same height  $h$  determined by  $N$  and involve same interactions of water molecules with graphite. Note that the extended meniscus is driven by interactions with one graphite surface, which should result in a somewhat smaller  $\Pi$  with respect to the modelled internal meniscus. This should lead to better agreement with our experimental results but possible corrections are neglected below. Water was supplied into the graphene capillaries from a relatively large reservoir placed on the left. The reservoir was terminated with a rigid graphene sheet that was allowed to move freely from left to right. Capillary pressure sucked water inside the channel and forced the sheet to move to the right. We applied a compensating force in the opposite direction to keep the sheet stationary. From the found force and the known cross-sectional area of the channel, the pressure  $P$  was calculated.

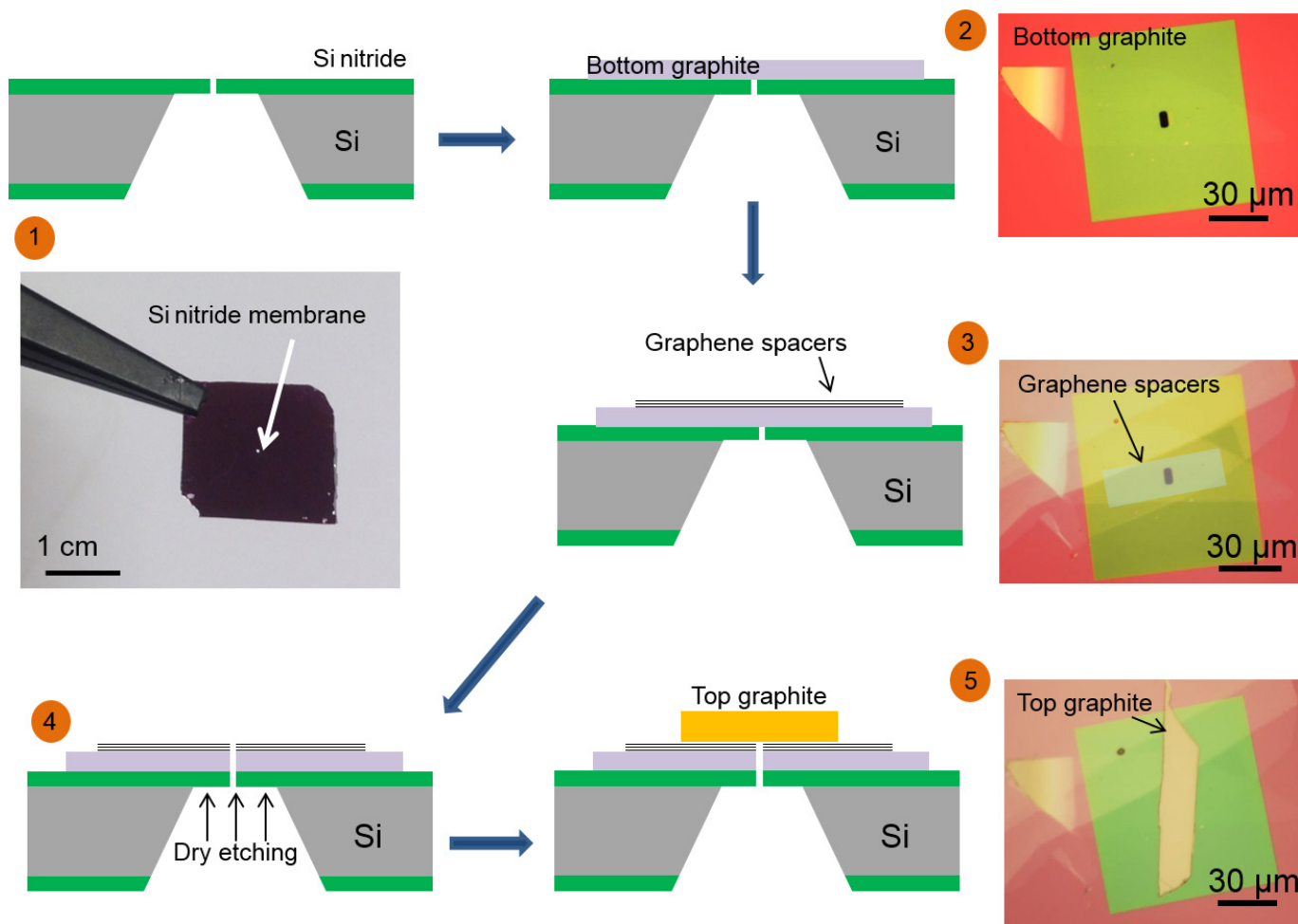
The results are shown in Extended Data Fig. 8b (solid symbols). The simulated capillary pressure rises notably faster than that expected from the classical term due to the curved meniscus (red curve). The steeper increase in  $P$  can be understood as due to the disjoining pressure  $\Pi$  that consists of several contributions, including the vdW pressure  $\Pi_{vdW}$  and entropic terms. The latter appear because of different densities of water inside and outside graphene nanocapillaries<sup>16,27</sup> as well as the enhanced structural order in nanoconfined water<sup>28,29,41,42,49,50</sup>. In our case, changes in  $\rho$  are relatively minor (inset of Extended Data Fig. 8b) leading to the corresponding entropic pressure<sup>16,27</sup> of  $< 50$  bar (magenta curve). Also,  $\Pi_{vdW} = A/(6\pi h^3)$  presents a relatively small effect, where  $A$  is the Hamaker constant for water–graphite interaction<sup>1,30</sup>. The  $\Pi_{vdW}$  contribution becomes notable only for  $N < 3$  because of the rapid  $h^{-3}$  dependence (blue curve). The total of the above three contributions is shown in Extended Data Fig. 8b by the green dashed curve. The remaining difference with respect to the MD-simulated dependence can be attributed to the entropic pressure due to the increased structural order in nanoconfined water<sup>27–31</sup>.

**Intrinsic collapse of monolayer capillaries.** To understand the complete blockage observed for all our devices with  $N = 1$ , we performed the following MD simulations. Graphene capillaries were modelled as flexible graphene layers stacked on top of each other with the interlayer distance  $a$ . One or two graphene layers were partially removed in the middle to create channels of 20 nm in width



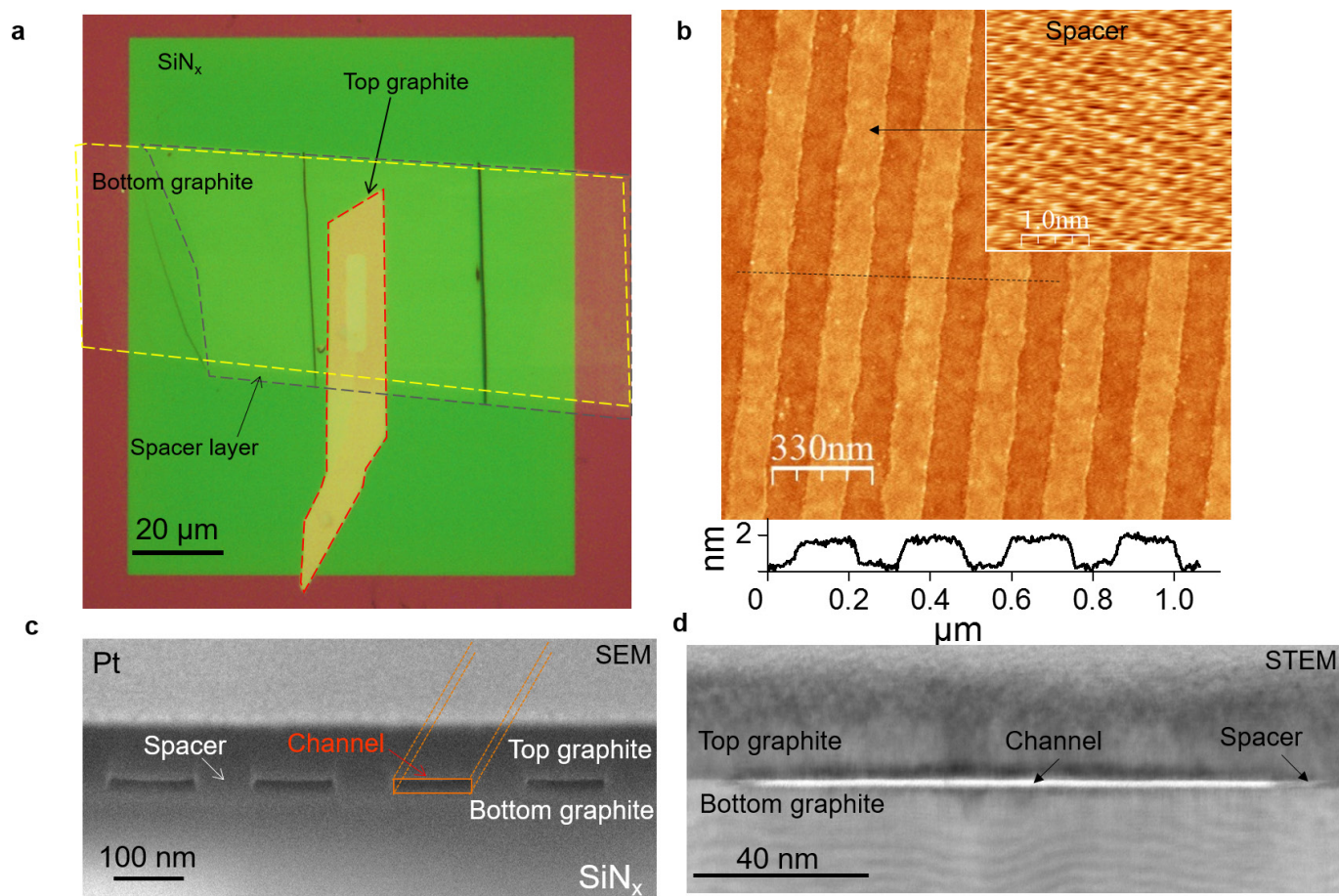
(Extended Data Fig. 9). We found that the walls of monolayer channels sagged already after several ps, independent of the thickness of graphite walls. In stark contrast, bilayer channels remained open. This behaviour is attributed to vdW attraction between capillary walls, which is sufficiently strong at short distances to deform the graphite bulk but rapidly vanishes with increasing the separation<sup>51</sup>.

32. Hu, S. *et al.* Proton transport through one-atom-thick crystals. *Nature* **516**, 227–230 (2014).
33. Withers, F. *et al.* Light-emitting diodes by band-structure engineering in van der Waals heterostructures. *Nat. Mater.* **14**, 301–306 (2015).
34. Schaffer, M., Schaffer, B. & Ramasse, Q. Sample preparation for atomic-resolution stem at low voltages by FIB. *Ultramicroscopy* **114**, 62–71 (2012).
35. Livesey, R. G. *Foundations of Vacuum Science and Technology* (Wiley & Sons, 1998).
36. Stein, D., Kruihof, M. & Dekker, C. Surface-charge-governed ion transport in nanofluidic channels. *Phys. Rev. Lett.* **93**, 035901 (2004).
37. Bocquet, L. & Charlaix, E. Nanofluidics, from bulk to interfaces. *Chem. Soc. Rev.* **39**, 1073–1095 (2010).
38. Pang, P., He, J., Park, J. H., Krstić, P. S. & Lindsay, S. Origin of giant ionic currents in carbon nanotube channels. *ACS Nano* **5**, 7277–7283 (2011).
39. Cussler, E. L. *Diffusion: Mass Transfer in Fluid Systems* 3rd edn (Cambridge Univ. Press, 2009).
40. Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
41. Hirunsit, P. & Balbuena, P. B. Effects of confinement on water structure and dynamics: a molecular simulation study. *J. Phys. Chem. C* **111**, 1709–1715 (2007).
42. Koga, K. & Tanaka, H. Phase diagram of water between hydrophobic surfaces. *J. Chem. Phys.* **122**, 104711 (2005).
43. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
44. Kannam, K. S., Todd, B. D., Hansen, J. S. & Daivis, P. J. Slip length of water on graphene: limitations of non-equilibrium molecular dynamics simulations. *J. Chem. Phys.* **136**, 024705 (2012).
45. Bocquet, L. & Barrat, J.-L. Flow boundary conditions from nano- to micro-scales. *Soft Matter* **3**, 685–693 (2007).
46. Neek-Amal, M., Peeters, F. M., Grigorieva, I. V. & Geim, A. K. Commensurability effects in the viscosity of nanoconfined water. *ACS Nano* **10**, 3685–3692 (2016).
47. Abascal, J. L. F. & Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys.* **123**, 234505 (2005).
48. Werder, T., Walther, J. H., Jaffe, R. L., Halicioglu, T. & Koumoutsakos, P. On the water–carbon interaction for use in molecular dynamics simulations of graphite and carbon nanotubes. *J. Phys. Chem. B* **107**, 1345–1352 (2003).
49. Kumar, P., Buldyrev, S. V., Starr, F. W., Giovambattista, N. & Stanley, H. E. Thermodynamics, structure, and dynamics of water confined between hydrophobic plates. *Phys. Rev. E* **72**, 051503 (2005).
50. Mosaddeghi, H., Alavi, S., Kowsari, M. H. & Najafi, B. Simulations of structural and dynamic anisotropy in nano-confined water between parallel graphite plates. *J. Chem. Phys.* **137**, 184703 (2012).
51. Lukas, M. *et al.* Catalytic subsurface etching of nanoscale channels in graphite. *Nat. Commun.* **4**, 1379 (2013).
52. White, F. M. *Viscous Fluid Flow* 2nd edn (McGraw-Hill, 1991).
53. Li, J.-L. *et al.* Use of dielectric functions in the theory of dispersion forces. *Phys. Rev. B* **71**, 235412 (2005).



**Extended Data Figure 1 | Microfabrication process flow.** (1) A micrometre-scale hole is prepared in a silicon nitride membrane. (2) Bottom graphite is transferred to cover the opening. (3) An array of graphene spacers is transferred on top. (4) The hole is extended into the graphite-graphene stack by dry etching. (5) Top graphite crystal is

transferred to cover the resulting aperture. The accompanying optical images (in natural colours) illustrate the results after each step for one of our devices. Graphene spacers are invisible in the photos and indicated by an opaque rectangle in (3). Steps 3 and 4 were often interchanged.

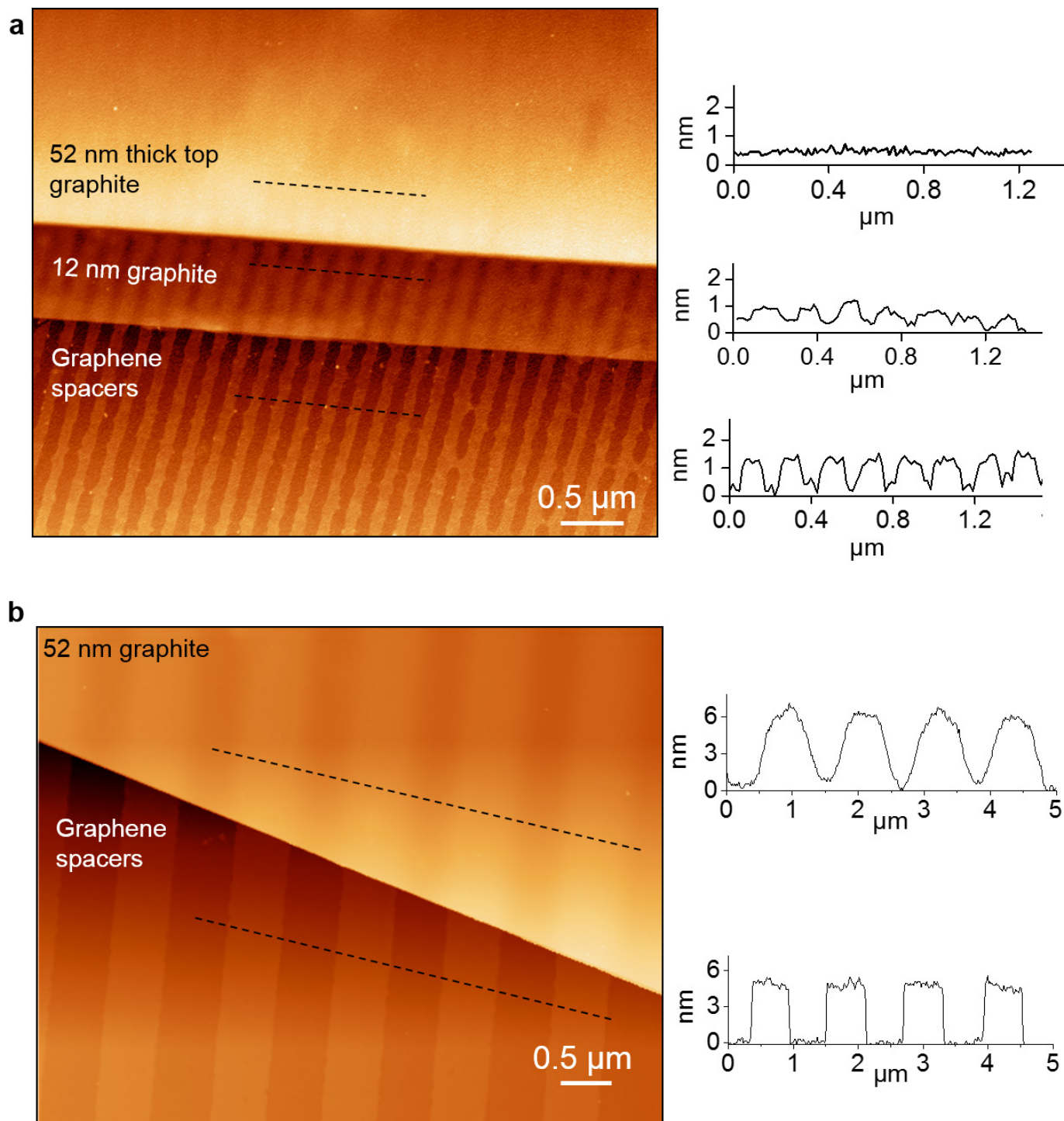


#### Extended Data Figure 2 | Additional images of graphene capillaries.

**a**, Optical image of a final device. The green region is the free-standing silicon nitride membrane. The Si wafer is seen in brown and the top graphite crystal (arrowed) in yellow. Red, yellow and grey contours indicate positions of the top graphite, bottom graphite and graphene spacers, respectively. The nearly-vertical dark lines are wrinkles in the bottom layer. **b**, AFM image of four-layer graphene spacers on top of a bottom graphite crystal (height profile along the dashed line is shown below the image). Inset, high-resolution scan (friction mode) from the region indicated by the arrow. The observation of the atomic lattice

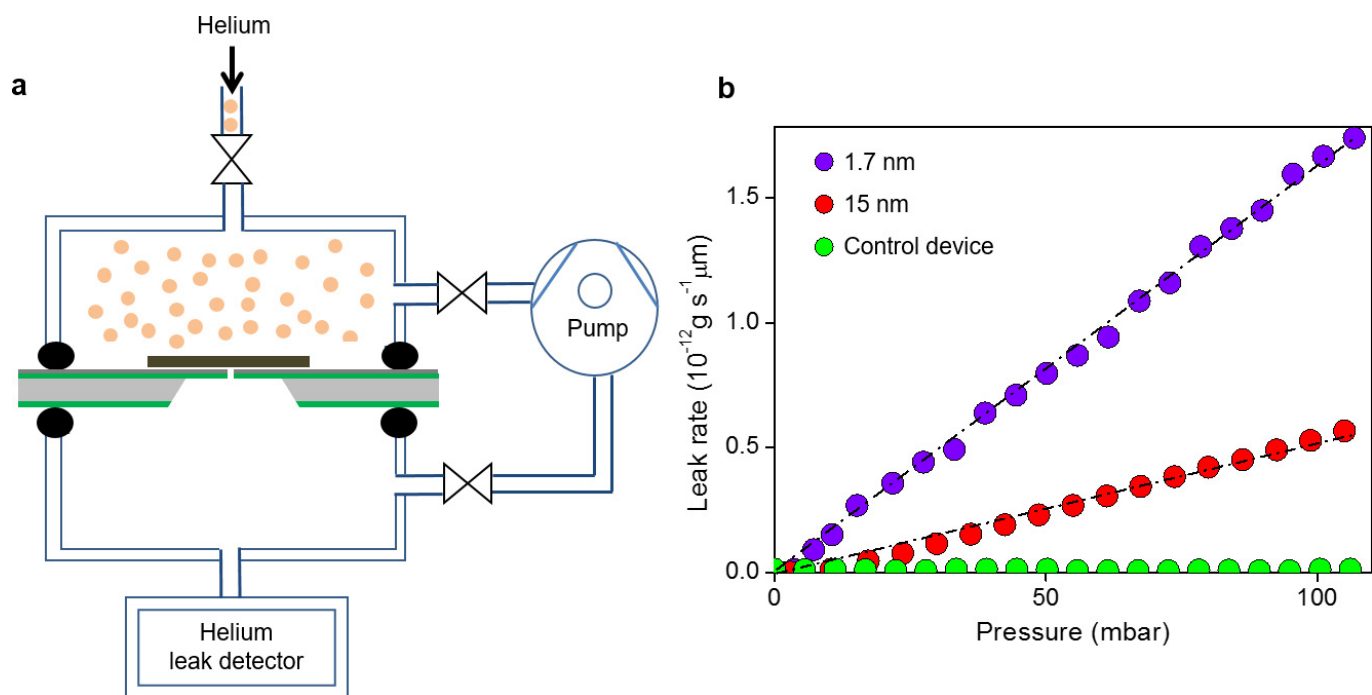
confirms that our assemblies have atomically smooth surfaces. Such smoothness is impossible to achieve using conventional materials and processes that invariably lead to the surface roughness exceeding the scale given by few-layer graphene spacers. Although the side walls of our channels are rough due to limitations of electron-beam lithography, we estimate that, because of the large ratios  $w/h$ , the side wall contribution to the flow resistance cannot exceed 5% even for our 10 nm devices<sup>52</sup>. **c**, SEM micrograph of a capillary device with  $h \approx 15$  nm. **d**, Bright field STEM image of a graphene capillary with  $N = 4$ . Spacers and channels are arrowed in **c** and **d**.





**Extended Data Figure 3 | Sagging of top graphite.** **a**, Left, AFM image of trilayer channels, which are covered by a graphite layer of varying thickness. **b**, Left, partial sagging of the top graphite into wide channels.

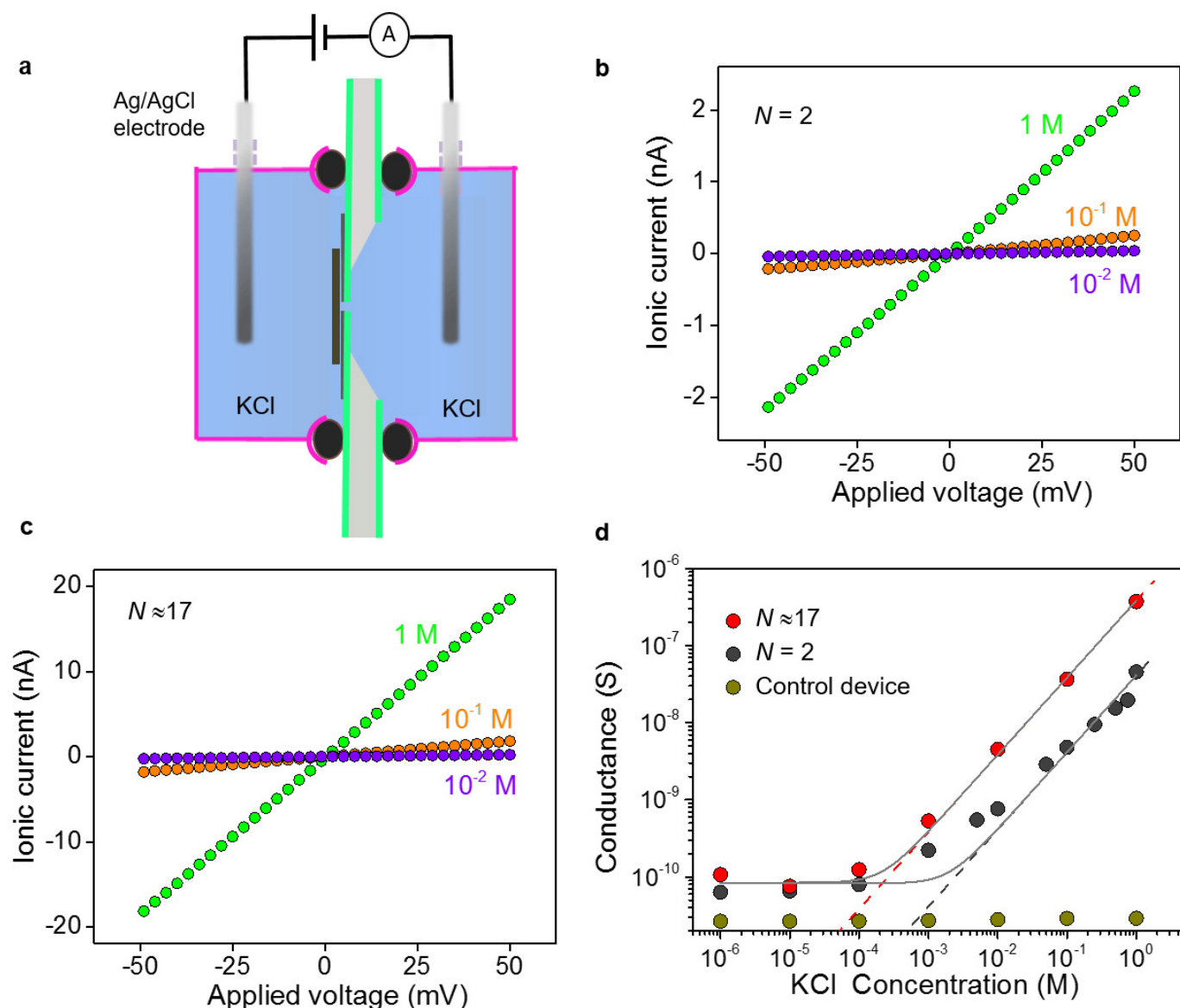
We can see that the top graphite bends down into the channels over their entire height  $h \approx 5$  nm. Right, height profiles that correspond to the traces shown by the dashed lines in the AFM images at left.



#### Extended Data Figure 4 | He leak through graphene capillaries.

**a**, Schematic of our set-up. Two vacuum chambers are separated by the silicon nitride wafer incorporating a nanocapillary device. Valves connect the chambers to a pump, a He leak detector and a gas inlet. He atoms are

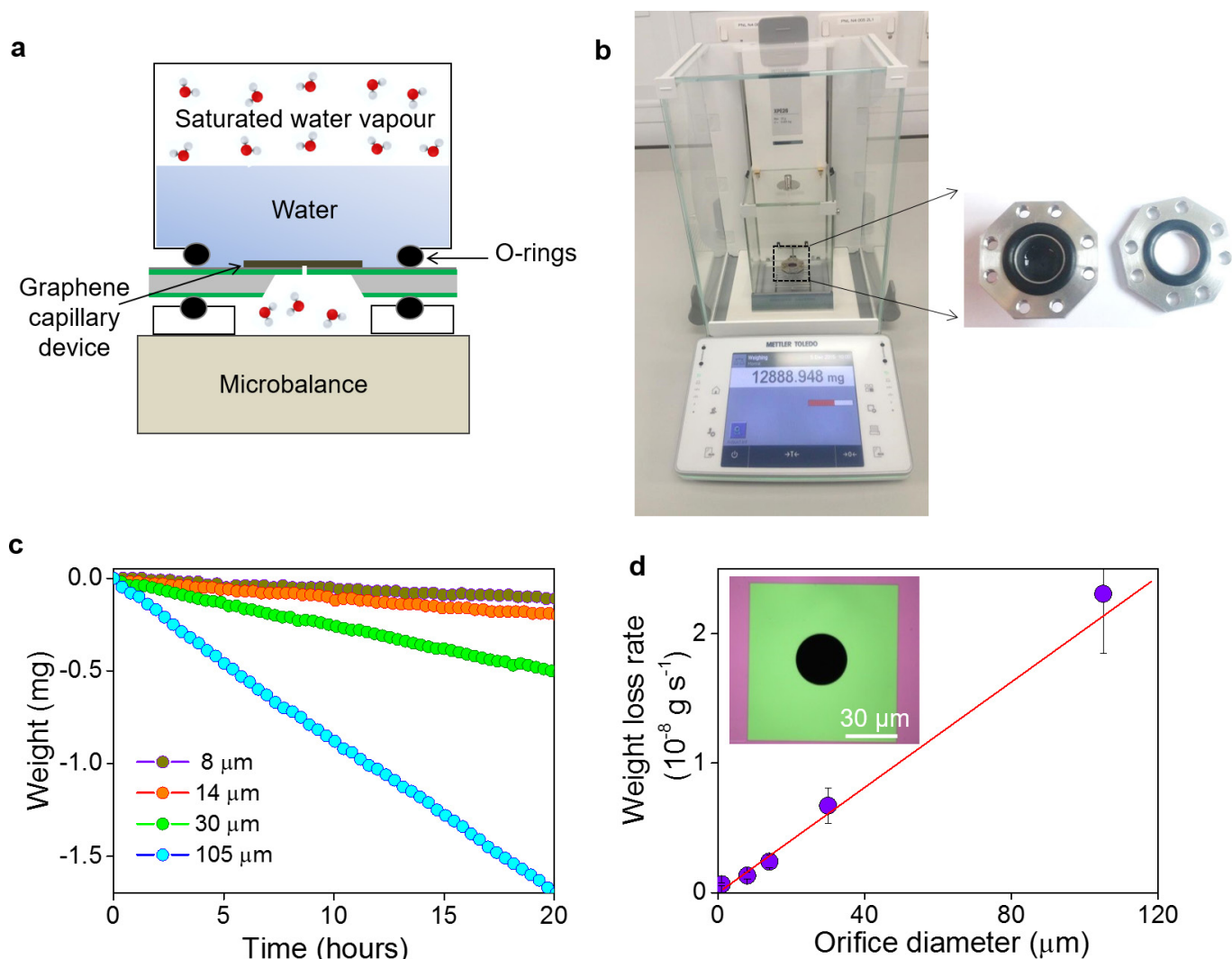
represented by filled orange circles. **b**, Leak rates normalized for  $1 \mu\text{m}$  length and given per channel as a function of applied pressure for capillary devices with  $N = 5$  and  $N \approx 45$  ( $h \approx 1.7 \text{ nm}$  and  $15 \text{ nm}$ , respectively), and a control device without graphene spacers ( $N = 0$ ).



**Extended Data Figure 5 | Ion transport through graphene nanochannels.** **a**, Schematic of our measurement set-up. A nanocapillary device fabricated on top of a Si nitride wafer (SiN is shown in green) is clamped using O-rings (black) to separate two containers (indicated by magenta lines). The containers are filled with a KCl solution (blue), and silver chloride-silver wires (dark grey) are used as electrodes to measure ionic conductance. **b**, Examples of current-voltage ( $I$ - $V$ ) characteristics of the smallest capillary devices ( $N=2$ ) at different KCl concentrations (labelling the curves;  $L$  ranges from  $2.8\ \mu\text{m}$  to  $7\ \mu\text{m}$ ). **c**, Same as **b** but for a device with  $N \approx 17$  of approximately the same average length  $\tilde{L}$  ( $L$  from

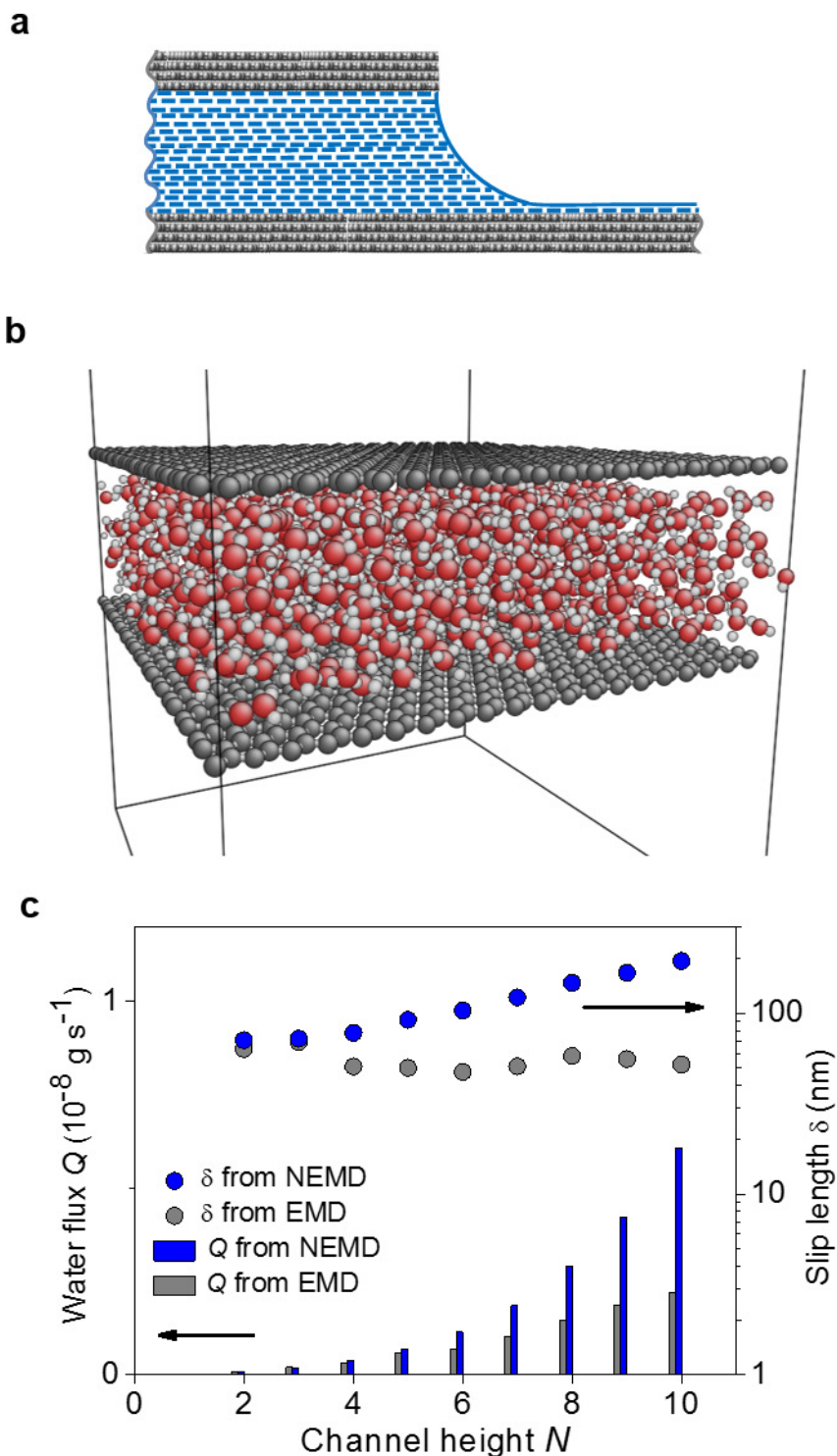
$1.7\ \mu\text{m}$  to  $7.3\ \mu\text{m}$ ). **d**, Ionic conductance for these devices as a function of KCl concentration,  $C$  (without normalizing for their slightly different  $\tilde{L}$ ). Both blank Si nitride wafers separating the reservoirs and control devices with  $N=0$  (no spacers but otherwise prepared using the same fabrication procedures) exhibited leakage conductance of the order of  $20\ \text{pS}$ , which did not change with  $C$  (olive symbols). The dashed lines show ionic conductance  $G$  expected from the bulk conductivity of KCl for the given channel dimensions. The solid curves are fits taking into account an additional parallel conductance due to the surface charge.





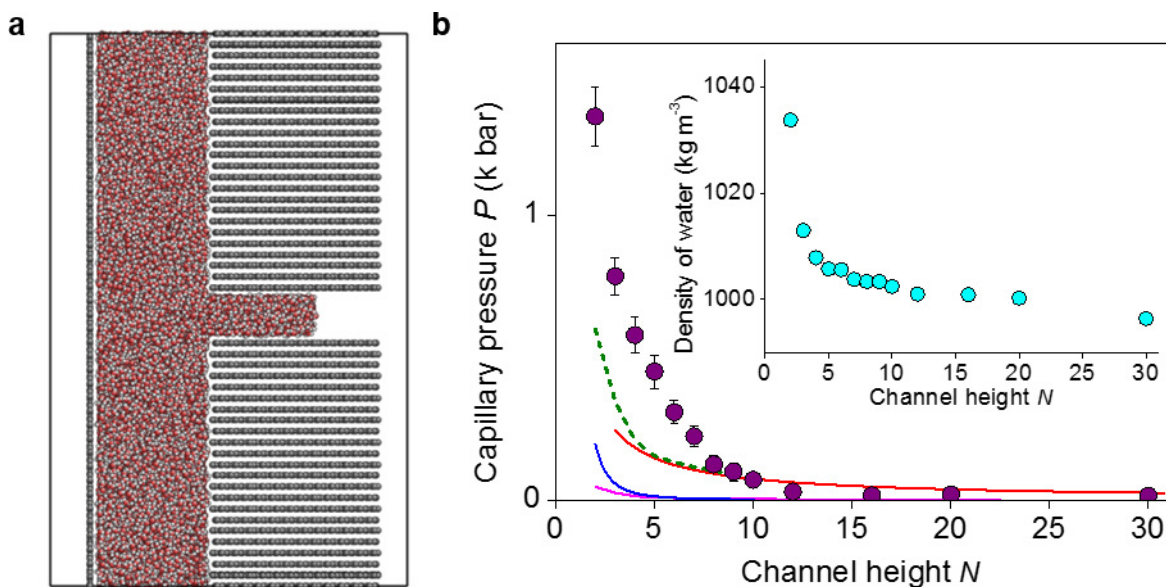
**Extended Data Figure 6 | Gravimetric measurements.** **a**, Extended schematic of the experimental set-up. A small aluminium container filled with water is sealed with a Si nitride wafer containing a graphene capillary device (total weight should not exceed  $\sim 15$  g to allow the required measurement accuracy). The container was weighed either upside down (water in contact with capillaries as shown in the sketch) or in the upright position as shown in the inset of Fig. 2a (capillaries are exposed to 100% RH). Both orientations resulted in the same  $Q$ . **b**, Photographs

of our gravimetric set-up. Main image, microbalance with our miniature container being weighed (its position is indicated by a dashed square). Image to the right, the container is open and the Si nitride wafer (that is clamped between the O-rings during measurements) is removed. **c**, Examples of water evaporation through apertures of different diameters,  $D$  (colour coded) **d**, Dependence of the evaporation rate on  $D$  (error bars, s.d.). Red line, best linear fit. Inset, optical micrograph (natural colour) of an aperture of  $30\ \mu\text{m}$  diameter, which is etched in a Si nitride membrane.



**Extended Data Figure 7 | Molecular dynamics simulations of water flow through graphene slits.** **a**, Our capillaries are filled with water and the driving pressure is determined by evaporation of the extended meniscus that appears at the capillary mouth. The meniscus is sketched in the drawing, showing a thin film of water propagating along the graphite surface<sup>26</sup>. **b**, MD set-up with the simulation box indicated by the black lines. The particular snapshot is for  $N=4$ . Dark grey balls represent

carbon atoms arranged into graphene planes. Red and light grey spheres show oxygen and hydrogen atoms of water molecules. **c**, Simulated slip length  $\delta$  and water flux  $Q$  as a function of  $N$ . In the case of NEMD,  $\delta$  (blue symbols) was calculated from the simulated  $Q$  (blue bars) using equation (1). Using  $\delta$  found from the EMD simulations (grey symbols) and the pressure gradient of  $10^{15} \text{ Pa m}^{-1}$ , equation (1) yields  $Q$  shown by the grey bars.

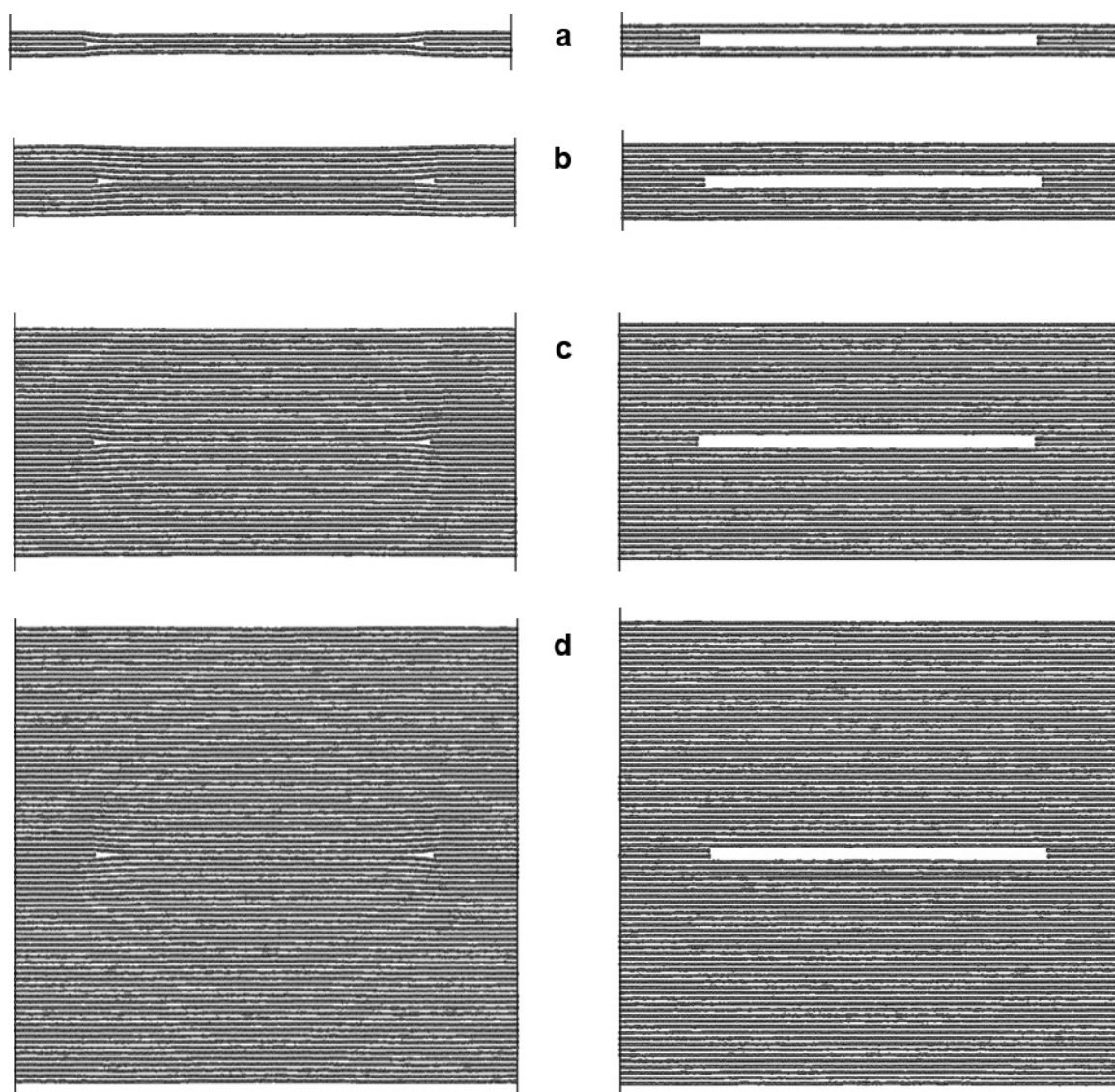


**Extended Data Figure 8 | MD simulations of capillary pressure.**

**a.** Our MD set-up for  $N=4$ . Colour coding as in Extended Data Fig. 7b. Graphene planes to the right represent a graphite crystal with four atomic planes removed. The vertical graphene plane is used as a movable membrane to apply a compensating force to stop the water meniscus from propagating to the right. **b.** Main figure, simulated capillary pressures (symbols with s.d. error bars). The red curve shows the best fit for large

$N$  using  $P_0 = 2\sigma\cos(\phi)/h$ , which yields  $\phi \approx 80^\circ$ . Blue and magenta curves show  $\Pi_{\text{vdW}}$  with the Hamaker constant  $A \approx 115$  zJ (ref. 53), and the entropic pressure due to changes in  $\rho$ , respectively. Dashed green curve, combined pressure from the three contributions. Inset, simulated density  $\rho$  of water confined between graphene sheets under external pressure of 1 bar.





**Extended Data Figure 9 | Micromechanical stability of graphene cavities.** Shown are snapshots of mono- and bilayer capillaries (left and right columns, respectively) after 100 ps of MD simulations.

**a–d,** Capillaries with different thicknesses of graphite walls, respectively 2, 6, 20 and 40 graphene layers.

# Evolution of global temperature over the past two million years

Carolyn W. Snyder<sup>1</sup>

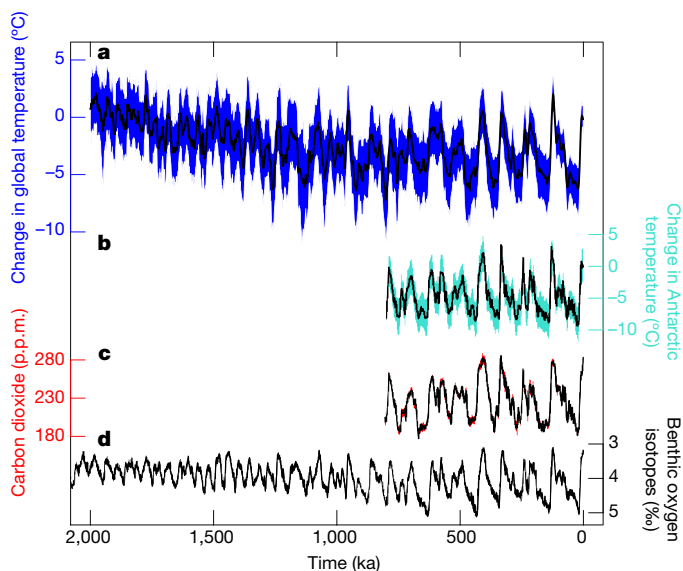
Reconstructions of Earth's past climate strongly influence our understanding of the dynamics and sensitivity of the climate system. Yet global temperature has been reconstructed for only a few isolated windows of time<sup>1,2</sup>, and continuous reconstructions across glacial cycles remain elusive. Here I present a spatially weighted proxy reconstruction of global temperature over the past 2 million years estimated from a multi-proxy database of over 20,000 sea surface temperature point reconstructions. Global temperature gradually cooled until roughly 1.2 million years ago and cooling then stalled until the present. The cooling trend probably stalled before the beginning of the mid-Pleistocene transition<sup>3</sup>, and predated the increase in the maximum size of ice sheets around 0.9 million years ago<sup>4–6</sup>. Thus, global cooling may have been a precondition for, but probably is not the sole causal mechanism of, the shift to quasi-100,000-year glacial cycles at the mid-Pleistocene transition. Over the past 800,000 years, polar amplification (the amplification of temperature change at the poles relative to global temperature change) has been stable over time, and global temperature and atmospheric greenhouse gas concentrations have been closely coupled across glacial cycles. A comparison of the new temperature reconstruction with radiative forcing from greenhouse gases estimates an Earth system sensitivity of 9 degrees Celsius (range 7 to 13 degrees Celsius, 95 per cent credible interval) change in global average surface temperature per doubling of atmospheric carbon dioxide over millennium timescales. This result suggests that stabilization at today's greenhouse gas levels may already commit Earth to an eventual total warming of 5 degrees Celsius (range 3 to 7 degrees Celsius, 95 per cent credible interval) over the next few millennia as ice sheets, vegetation and atmospheric dust continue to respond to global warming.

Reconstructions of several key climate variables are available with high temporal resolution across past glacial cycles, such as polar temperature, atmospheric greenhouse gas (GHG) concentrations, sea surface temperature (SST), deep-water temperature (DWT) and sea level (see, for example, Extended Data Tables 1–3). Yet global average surface temperature (GAST) has been reconstructed for only a few isolated windows of time<sup>1,2</sup>, and continuous reconstructions across glacial cycles remain elusive. The lack of continuous GAST reconstructions has constrained model–data comparisons to particular extreme points in time, such as the Last Glacial Maximum (LGM), but multiple time points are critical for characterizing the uncertainty in relationships estimated from palaeoclimate reconstructions<sup>7,8</sup>. The potential power of a continuous GAST record has been recently demonstrated<sup>8</sup>; GAST was reconstructed for the past 22,000 years (kyr) and used to clarify carbon dioxide's role in driving global climate change across glacial cycles. The present research creates a continuous record of GAST across a much longer timescale.

Previous continuous reconstructions of GAST across glacial cycles used only a single proxy record that was scaled linearly<sup>2,4,9–12</sup> or modelled<sup>13</sup> to estimate global values. This Letter presents a spatially weighted proxy reconstruction of GAST over the past 2 million years

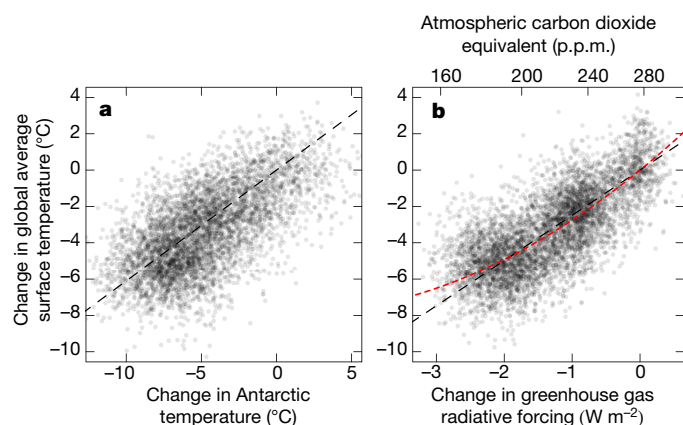
(Fig. 1a), estimated using a multi-proxy database compilation of over 20,000 SST point reconstructions from 59 ocean sediment cores (Extended Data Tables 1, 2). This research uses probabilistic simulations across multiple sources of uncertainty to estimate credible intervals at 1-kyr intervals, and validates the new reconstruction against previous estimates. The new GAST reconstruction can provide key insights into several major palaeoclimate questions, including the magnitude and stability of polar amplification, the state dependence of Earth system sensitivity (ESS, see below), and the role of global temperature in the mid-Pleistocene transition (MPT).

A comparison of GAST to Antarctic temperature reconstructions for the past 800 kyr finds that GAST and Antarctic temperature<sup>14</sup> are closely coupled across glacial cycles with a correlation of 0.72 (0.59–0.81, 95% credible interval, hereafter 'interval')—a high correlation given that the GAST reconstruction is estimated independently of the ice core records. There is a linear relationship of 0.61 °C (0.43–0.85 °C, 95% interval) change in GAST for every 1 °C change in Antarctic temperature (Fig. 2a) that does not significantly change over the past 800 kyr (Extended Data Fig. 6a). Some previous research on climate sensitivity over the past 800 kyr has assumed that changes in GAST are similar to half the magnitude of changes in Antarctic temperature<sup>9,12,15</sup>. On the basis of the new GAST reconstruction, there is an 87%



**Figure 1 | Reconstruction of global average surface temperature (GAST) over the past 2 million years compared to other key palaeoclimate variables.** **a**, GAST as temperature deviation (in °C) from present (average over 0–5 ka) in blue. **b**, Stacked reconstruction of change in Antarctic temperature<sup>14</sup> (°C) in cyan. **c**, Stacked reconstruction of atmospheric CO<sub>2</sub> concentrations<sup>18</sup> (p.p.m.) in red. **d**, Stack of deep-sea oxygen isotopes<sup>30</sup>, δ<sup>18</sup>O (‰), in grey. In all panels, the solid black lines show the median estimate and the colour shaded areas show the 95% interval.

<sup>1</sup>Interdisciplinary Program in Environment and Resources, Stanford University, Stanford, California 94305, USA.



**Figure 2 | Relationship of changes in GAST to changes in Antarctic temperature and GHG radiative forcing over the past 800 kyr.** **a, b,** Each point represents randomly sampled estimates from simulations of GAST plotted against Antarctic temperature<sup>14</sup> (**a**) and GHG radiative forcing<sup>14,17,18</sup> (**b**) over the past 800 kyr. The dashed black line shows the median estimated relationship in °C per °C in **a** and in °C per  $\text{W m}^{-2}$  in **b**. The red dashed line shows the median estimated quadratic relationship in **b**.

probability that such an assumption underestimates global temperature and thus climate sensitivity.

Polar amplification can be estimated as change in Antarctic temperature for every 1 °C change in GAST, here estimated as 1.6 °C per °C (1.2–2.3 °C per °C, 95% interval). Estimates of polar amplification are complicated because the elevation of ice sheets changes during glacial cycles due to changes in accumulation and isostasy<sup>9,14,16</sup>, and current ice sheet glaciological models disagree markedly with the ice topography used in LGM model simulations for Antarctica<sup>9</sup>. Climate models estimate that polar amplification of uncorrected Antarctic temperature will be nonlinear and lower in colder states, ~2 °C per °C for the LGM and ~1.2 °C per °C for future warming<sup>9,16</sup>, but ~1.2 °C per °C for both when Antarctic temperature is elevation-corrected<sup>16</sup>. The present research uses a reconstruction of Antarctic temperature that is elevation-corrected<sup>14</sup>, but corrections are highly dependent upon uncertain ice sheet assumptions<sup>14,16</sup>. A comparison of the new GAST reconstruction with Antarctic temperature finds a quadratic relationship to not be significant as predicted by the models (Fig. 2a), suggesting the elevation correction of Antarctic temperature is adequate. However, the magnitude of the polar amplification estimate from the new GAST reconstruction is significantly higher than predicted by many models (97.5% probability above 1.2)<sup>9,16</sup>. It is worth noting that those same models underestimate elevation-corrected Antarctic temperature change at the LGM by a factor of 1.2–2.3 (ref. 16).

Other research has assumed that changes in DWT can be used as a direct proxy for GAST<sup>4,9,10</sup> or doubled to estimate GAST<sup>11</sup>. A comparison between GAST and 12 DWT reconstructions from three different methods<sup>4–6</sup> finds highly variable results, with median correlations varying between 0.3 and 0.8 and median linear relationships varying between 1.4 °C and 3.5 °C change in GAST per 1 °C change in DWT (Extended Data Table 3). The observed attenuation of the global temperature signal and the reduced correlation may be caused by deep-water cooling being limited by water's freezing temperature and/or by changes in ocean circulation<sup>4,6,17</sup>. These results demonstrate the high uncertainty in inferring GAST from any single DWT reconstruction.

High-resolution estimates of atmospheric GHG concentrations are also available from Antarctic ice core records<sup>14,18</sup> over the past 800 kyr. The GAST reconstruction reveals a remarkably stable relationship between GAST and GHG radiative forcing<sup>14,17,18</sup> over the past 800 kyr with a correlation of 0.82 (0.66–0.92, 95% interval), stronger than the correlation between GAST and Antarctic temperature (Fig. 2b).

The concept of  $S_{\text{[GHG]}}$  has been defined<sup>2</sup> as the total global climate response over millennial timescales from changes in ice sheets, dust and vegetation, as well as from the feedbacks included in 'equilibrium climate sensitivity'—water vapour, lapse rate, sea ice, snow cover, clouds and ocean heat uptake, but does not include carbon cycle feedbacks<sup>1,2</sup>. A comparison of the new GAST reconstruction with GHG radiative forcing estimates  $S_{\text{[GHG]}}$  as 2.5 °C (1.8–3.6 °C, 95% interval) change in GAST per  $1 \text{ W m}^{-2}$  change in GHG radiative forcing (Fig. 2b), and finds that the relationship does not change significantly over the past 800 kyr (Extended Data Fig. 6b). This  $S_{\text{[GHG]}}$  estimate translates to a 9 °C (7–13 °C, 95% interval) change in GAST per doubling of atmospheric carbon dioxide ( $3.7 \text{ W m}^{-2}$ ), which has often been called ESS<sup>2,19,20</sup>.

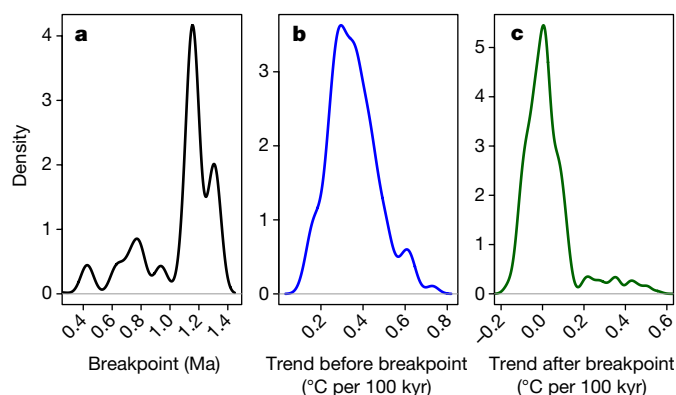
Attenuation of the  $S_{\text{[GHG]}}$  relationship is apparent in deep glacial states and a quadratic relationship is found to be a significantly better fit than a linear relationship (Fig. 2b). However, it is unclear whether such a quadratic relationship would apply in warmer states—when the bottom quarter of the record is removed, a quadratic relationship is not significant. Previous research also found  $S_{\text{[GHG]}}$  to be climate state dependent, and most studies find higher values for the late Quaternary than for the Pliocene<sup>2</sup>. The presence of large ice sheets in the Quaternary is probably a major cause. Yet little research has focused on the potential variation of  $S_{\text{[GHG]}}$  within the late Quaternary. Masson-Delmotte *et al.*<sup>9</sup> also found attenuation within the late Quaternary in deep glacial states, estimating a parabolic relationship. The observed attenuation of late Quaternary  $S_{\text{[GHG]}}$  seems to suggest there is a limit to the power of positive climate feedbacks, such as from sea ice and ice sheets, as ice sheet size increases in deep glacial states.

Because  $S_{\text{[GHG]}}$  and ESS are climate state dependent, it is most useful to compare this result to other estimates from the late Quaternary<sup>2</sup>. Rohling *et al.*<sup>2</sup> found a similar ESS estimate of 8.5 °C, but did not include a comparable probabilistic analysis in their estimate. Hansen *et al.*<sup>11,15</sup> both estimated ESS of 6 °C, assuming GAST is half the Antarctic temperature change or twice the DWT change, respectively. The present research finds that there is a 99% probability that ESS for the late Quaternary is higher than 6 °C.

The new GAST reconstruction also can provide insight into the MPT. The causes of the MPT, when the Earth's climate shifted from glacial cycles with periods of about 41 kyr to those with quasi-100-kyr periods, are not well understood and debates continue on the potential linkage between different orbital changes and the quasi-100-kyr cycles<sup>21–24</sup>. Some theories explain the MPT with changes in nonlinear feedbacks internal to the climate system, such as changes to ice sheets, sea ice or ocean circulation<sup>3,21,22,25–28</sup>. An alternative theory is the erosion of continental regolith underneath the ice sheets enabling the growth of thicker ice sheets<sup>3</sup>.

Probabilistic breakpoint analysis is used to identify any changes in cooling trends across the past 2 Myr in the new GAST reconstruction, as well as the timing of the cooling trend changes. Such analysis find a strong cooling trend after 2 Myr ago (Ma) that then stops most probably at 1.2 Ma (median estimate), with a 72% probability that GAST cooling stopped by 1.1 Ma and 77% by 0.9 Ma (Fig. 3a). The timing of when the global cooling trend stops roughly corresponds to estimates of the beginning of the broad MPT, which is estimated to occur over the general period of 1.25 to 0.7 Ma based on spectral analysis of oxygen isotopes<sup>3</sup>. Before roughly 1.2 Ma, global temperature cooled gradually by approximately 0.34 °C (0.16–0.62 °C, 95% interval) per 100 kyr (Fig. 3b). However, since 1.2 Ma, GAST stabilized with no significant change in global temperature, 0.007 °C (–0.12 to 0.42 °C, 95% interval) per 100 kyr (Fig. 3c). From 1.2 to 0.5 Ma, the behaviour of GAST across glacial–interglacial cycles gradually shifted to quasi-100-kyr cycles with larger amplitudes of change, as seen in Fig. 1a. Although average GAST did not continue to cool after roughly 1.2 Ma, GAST did show a particularly large amplitude for the glacial cycle at 0.9 Ma of 7 °C (4–10 °C, 95% interval), which is similar in magnitude to more recent, post-MPT glacial cycles. These findings of gradual cooling probably pre-dating the MPT and a gradual shift to quasi-100-kyr





**Figure 3 | Probabilistic breakpoint analysis of global temperature trends over the past 2 million years.** Shown are empirically fitted frequency distributions for the timing of when the trend in global temperature changes (a; in black), the global temperature trend before the breakpoint (b; in blue), and the global temperature trend after the breakpoint (c; in green).

cycles are consistent with some previous SST and DWT research<sup>6,25,29</sup>. The global cooling trend also is synchronous with the development of the equatorial Pacific cold tongue and bipolar cooling estimated from ~1.8 Ma to ~1.2 Ma (ref. 29). However, GAST does not exhibit the intensified cooling across the MPT seen in some individual SST records and predicted by some MPT theories<sup>25</sup>.

Several MPT theories employ increases in ice sheet size to explain the change in nonlinear climate feedbacks at the MPT, and hypothesize that global cooling could be the causal mechanism for such ice sheet growth<sup>21,22,26,27</sup>. Analyses of orbital responses across the MPT similarly suggested that global cooling could have enabled the skipping of obliquity cycles<sup>22,23</sup>. The present research provides evidence of such global cooling before the MPT. However, the global cooling probably pre-dates the rapid ice sheet growth observed at the deep glacial period around 0.9 Ma (refs 4–6) and the development of the first quasi-100-kyr cycle by 300 kyr ago (ka). Thus, either additional explanation is required to explain the lag after global cooling before MPT climate changes or the MPT changes may have been caused by a mechanism not linked to global temperature, such as erosion of continental regolith<sup>3</sup> or orbital changes without internal climate changes<sup>24</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 27 January; accepted 24 August 2016.**

**Published online 26 September 2016.**

1. Masson-Delmotte, V. *et al.* in *Climate Change 2013: The Physical Science Basis* Ch. 5 (eds Stocker, T. F. *et al.*) 383–464 (Cambridge Univ. Press, 2013).
2. Rohling, E. J. *et al.* Making sense of palaeoclimate sensitivity. *Nature* **491**, 683–691 (2012).
3. Clark, P. U. *et al.* The middle Pleistocene transition: characteristics, mechanisms, and implications for long-term changes in atmospheric pCO<sub>2</sub>. *Quat. Sci. Rev.* **25**, 3150–3184 (2006).
4. Elderfield, H. *et al.* Evolution of ocean temperature and ice volume through the Mid-Pleistocene climate transition. *Science* **337**, 704–709 (2012).
5. Rohling, E. J. *et al.* Sea-level and deep-sea-temperature variability over the past 5.3 million years. *Nature* **508**, 477–482 (2014).
6. Bates, S. L., Siddall, M. & Waelbroeck, C. Hydrographic variations in deep ocean temperature over the mid-Pleistocene transition. *Quat. Sci. Rev.* **88**, 147–158 (2014).
7. Rohling, E., Medina-Elizalde, M., Shepherd, J., Siddall, M. & Stanford, J. Sea surface and high-latitude temperature sensitivity to radiative forcing of climate over several glacial cycles. *J. Clim.* **25**, 1635–1656 (2012).

8. Shakun, J. D. *et al.* Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nature* **484**, 49–54 (2012).
9. Masson-Delmotte, V. *et al.* EPICA Dome C record of glacial and interglacial intensities. *Quat. Sci. Rev.* **29**, 113–128 (2010).
10. Zachos, J., Pagani, M., Sloan, L., Thomas, E. & Billups, K. Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
11. Hansen, J., Sato, M., Russell, G. & Kharecha, P. Climate sensitivity, sea level and atmospheric carbon dioxide. *Phil. Trans. R. Soc. Lond. A* **371**, (2013).
12. Chylek, P. & Lohmann, U. Aerosol radiative forcing and climate sensitivity deduced from the last glacial maximum to Holocene transition. *Geophys. Res. Lett.* **35**, L04804 (2008).
13. van de Wal, R. S. W., de Boer, B., Lourens, L. J., Kohler, P. & Bintanja, R. Reconstruction of a continuous high-resolution CO<sub>2</sub> record over the past 20 million years. *Clim. Past* **7**, 1459–1469 (2011).
14. Parrenin, F. *et al.* Synchronous change of atmospheric CO<sub>2</sub> and Antarctic temperature during the last deglacial warming. *Science* **339**, 1060–1063 (2013).
15. Hansen, J. *et al.* Target atmospheric CO<sub>2</sub>: where should humanity aim? *The Open Atmos. Sci. J.* **2**, 217–231 (2008).
16. Masson-Delmotte, V. *et al.* Past and future polar amplification of climate change: climate model intercomparisons and ice-core constraints. *Clim. Dyn.* **26**, 513–529 (2006).
17. Köhler, P. *et al.* What caused Earth's temperature variations during the last 800,000 years? Data-based evidence on radiative forcing and constraints on climate sensitivity. *Quat. Sci. Rev.* **29**, 129–145 (2010).
18. Bereiter, B. *et al.* Revision of the EPICA Dome C CO<sub>2</sub> record from 800 to 600kyr before present. *Geophys. Res. Lett.* **42**, 542–549 (2015).
19. Pagani, M., Liu, Z., LaRiviere, J. & Ravelo, A. C. High Earth-system climate sensitivity determined from Pliocene carbon dioxide concentrations. *Nat. Geosci.* **3**, 27–30 (2010).
20. Lunt, D. J. Earth system sensitivity inferred from Pliocene modelling and data. *Nat. Geosci.* **3**, 60–64 (2010).
21. Raymo, M. E. The timing of major climate terminations. *Paleoceanography* **12**, 577–585 (1997).
22. Huybers, P. Early Pleistocene glacial cycles and the integrated summer insolation forcing. *Science* **313**, 508–511 (2006).
23. Huybers, P. & Wunsch, C. A depth-derived Pleistocene age model: uncertainty estimates, sedimentation variability, and nonlinear climate change. *Paleoceanography* **19**, PA1028 (2004).
24. Imbrie, J. Z., Imbrie-Moore, A. & Lisiecki, L. E. A phase-space model for Pleistocene ice volume. *Earth Planet. Sci. Lett.* **307**, 94–102 (2011).
25. McClymont, E. L., Sosdian, S. M., Rosell-Melé, A. & Rosenthal, Y. Pleistocene sea-surface temperature evolution: early cooling, delayed glacial intensification, and implications for the mid-Pleistocene climate transition. *Earth Sci. Rev.* **123**, 173–193 (2013).
26. Tziperman, E. & Gildor, H. On the mid-Pleistocene transition to 100-kyr glacial cycles and the asymmetry between glaciation and deglaciation times. *Paleoceanography* **18**, PA000627 (2003).
27. Abe-Ouchi, A. *et al.* Insolation-driven 100,000-year glacial cycles and hysteresis of ice-sheet volume. *Nature* **500**, 190–193 (2013).
28. Ganopolski, A. & Calov, R. The role of orbital forcing, carbon dioxide and regolith in 100 kyr glacial cycles. *Clim. Past* **7**, 1415–1425 (2011).
29. Martínez-García, A., Rosell-Melé, A., McClymont, E. L., Gersonde, R. & Haug, G. H. Subpolar link to the emergence of the modern Equatorial Pacific Cold Tongue. *Science* **328**, 1550–1553 (2010).
30. Lisiecki, L. E. & Raymo, M. E. A Pliocene-Pleistocene stack of 57 globally distributed benthic  $\delta^{18}\text{O}$  records. *Paleoceanography* **20**, PA1003 (2005).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** I thank S. Schneider, C. Field, C. Tebaldi, R. Dunbar, K. Caldeira, C. Warshaw, H. Elderfield and R. Samworth for advice and feedback. I am indebted to many scientists for supplying their proxy data records (see Extended Data Tables 1–3), and to NOAA's National Centers for Environmental Information and PANGAEA. This study was supported by a National Science Foundation Graduate Research Fellowship. The views expressed in this article are those of the author and do not necessarily reflect the views or policies of the US Environmental Protection Agency.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The author declares no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.W.S. (carolyn.snyder@gmail.com).

**Reviewer Information** *Nature* thanks E. J. Rohling and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**GAST reconstruction. Overall approach.** This research estimates GAST from local SST proxy-based reconstructions through five steps. First, I collect SST proxy-based reconstructions and estimate proxy uncertainty from a literature review. Second, I interpolate the SST reconstructions to common 1-kyr intervals and estimate dating uncertainty. Third, I estimate average SST values for latitudinal zones using a variety of possible spatial weighting schemes. Fourth, I analyse Paleoclimate Modelling Intercomparison Project (PMIP) model simulations to obtain an estimate of the relationship between average SST over latitudinal zones to GAST. I use the estimated scalar to linearly scale the average changes in SST for latitudinal zones to changes in GAST. Last, I use several approaches to test the validity and sensitivity of this approach.

The biggest challenge in this research is that the primary continuous temperature reconstructions available over the past 2 Myr are mostly from SST proxy records. Available terrestrial temperature reconstructions are too infrequent and limited in spatial distribution to be used for a global reconstruction at this time. Thus, this research develops a method to estimate GAST from a collection of SST estimates. Such a scaling must address the fact that SST records do not adequately cover the entire Earth surface, and that by definition, SST records do not include records of temperature change over land. This is especially important given that temperature change is amplified over land and at the poles relative to the oceans. The goal of this research is to develop a method that is transparent in its assumptions and associated estimates of uncertainty and could be applied in the same way across the past 2 Myr, even though there are far fewer SST reconstructions available farther back in time. Because GAST estimates are often used to investigate questions related to climate sensitivity, the designed approach should not be dependent on any assumptions of climate sensitivity.

To investigate potential approaches to this challenge, this research uses the PMIP global climate models<sup>31,32</sup> because the PMIP model experiments provide estimates of local SST and air surface temperature as well as GAST for both the LGM and the pre-industrial state, thus covering most of the range of temperatures over the past 2 Myr. Because the highest latitude SST reconstructions are at about 60° N/S, I compared the ratio of change in GAST to the average change in SST in all SST grid cells 60° N–60° S. The estimated scaling factors are found to not be correlated with other model features, such as climate sensitivity and LGM GAST (Extended Data Fig. 4). Hargreaves *et al.*<sup>33</sup> similarly find that LGM GAST is not correlated with the model's climate sensitivity values. This is very important, since a main goal of this approach is to develop a method that is not linked to a particular climate model or a particular estimate of climate sensitivity. Another approach would be to scale just from average tropical SST changes, but that would ignore valuable data from high latitude records, would be more uncertain due to the relative larger contribution of the scalar value, and would be more model dependent.

It is important to note that the concept of a regional average SST over latitudes 60° N–60° S does not have direct physical meaning or relevance as it is not a regional average since it does not include the land area in latitudes 60° N–60° S and only averages over SST grid cells. Nor does the described scalar of change in GAST relative to change in SST over 60° N–60° S. However, taking a simple average over latitudinal bands is a stable and transparent way to summarize the available SST reconstructions. Moreover, there is no assumed functional relationship of SST with latitude or with GAST, except for the single coarse scaling metric. This method ensures that the full change in GAST at glacial maxima is captured in the final estimated reconstruction by using the LGM scalar from the PMIP models.

Rohling *et al.*<sup>7</sup> used a somewhat similar approach of a spatial average of SST to estimate global climate sensitivity. They estimated a quadratic relationship of change in SST over latitude from 36 SST reconstructions and integrated that function to calculate a global mean response of SST. They then adjusted for a stronger terrestrial response to scale to a larger estimate of global climate sensitivity, but they did not produce an estimate of GAST.

**SST proxy-based reconstructions.** This analysis utilizes a multi-proxy database compilation of all available and reliable SST reconstructions that cover at least the past 100 kyr. The SST database includes 61 SST proxy reconstructions from 59 ocean sediment cores: 29 using alkenone unsaturation indices ( $U_{37}^K$ ), 17 using ratios of Mg/Ca in planktonic foraminifera, and 16 based upon microfossil abundances (using transfer functions for planktonic foraminifera and radiolarians) (Extended Data Tables 1, 2; Extended Data Figs 1, 2; Supplementary Data). A multi-proxy approach enables a reduction of the uncertainties and potential biases specific to each proxy method by combining estimates from several independent proxies<sup>34</sup>.

Proxy methods have a variety of potential sources of error, including proxy measurement, seasonality, species dependence, productivity, water column depth, mixing, and dissolution and other post-depositional alteration. Estimates of the measurement and calibration errors are available from laboratory and

field experiments for the different proxy methods (for example, from alkenone indices<sup>35,36</sup>, Mg/Ca ratios<sup>37,38</sup> and species assemblages<sup>39</sup>) and typically range from 1 °C to 3 °C for two standard deviations. However, the published uncertainty estimates often do not include considerations of structural uncertainty from the assumptions of the proxy method. Therefore, I use the upper range of the published values of 3 °C (95% interval) as an estimate for the combined uncertainty for each of the SST proxy methods.

**Dating uncertainty and SST reconstruction interpolation.** It is imperative that comparisons between palaeoclimate records include the uncertainty in matching which parts of each record occurred at the same point in time<sup>40,41</sup>. To interpolate each SST reconstruction to a common 1-kyr timescale, I estimated a weighted average of the SST reconstruction for each time point on the 1-kyr timescale. The weights are based on the distance in time between the reconstruction value and the time point of interest. The bandwidth is set by the dating uncertainty for that time point. I use the published age scales for the SST reconstructions. I use the estimate of 10 kyr (95% interval) for dating uncertainty from orbital tuning<sup>23,41–43</sup>, unless papers provide specific estimates of uncertainty in their timescales. The uncertainty in the estimated interpolated value is estimated from a weighted average of the squared differences between the reconstruction values and the estimated interpolated value. I implement this method using a Nadaraya–Watson kernel-weighted local constant regression, using the function *ksmooth* in the R statistical program (<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/ksmooth.html>). This method results in larger SST uncertainty during periods of rapid change than during periods of stable SST, thus reflecting the varying potential impact of dating uncertainty.

**SST averages over a latitudinal zone.** The spatial distribution of the available SST records is too sparse to use spatial statistics to estimate average SST. An alternative approach is to apply a simple assumption of a quadratic change in SST with latitude<sup>7</sup>, but such a method adds a large amount of uncertainty when there are fewer available reconstructions. The simplest approach would be a direct average of all SST reconstructions equally, but that ignores the known general amplification of change in SST with latitude and would make the estimate very dependent on the particular distribution of the available reconstructions. The proposed method is a middle ground: SST is first averaged across records within a single latitudinal zone and then the latitudinal zones are summed using applicable spatial weights. Because it is spatially averaged, it is not purely driven by high latitude records, and the reduction in uncertainty from multiple records in a given latitudinal zone is captured.

To explore how the 60° N–60° S average SST estimate varies with different latitudinal zone boundaries, I use 9 different possible configurations of latitudinal zone boundaries used with equal weights in the final ensemble:

- Four zones:  
60° N–30° N, 30° N–0°, 0°–30° S, 30° S–60° S (equal degrees);  
60° N–25.7° N, 25.7° N–0°, 0°–25.7° S, 25.7° S–60° S (equal areas);  
60° N–20° N, 20° N–0°, 0°–20° S, 20° S–60° S;  
60° N–35° N, 35° N–0°, 0°–35° S, 35° S–60° S.
- Six zones:  
60° N–40° N, 40° N–20° N, 20° N–0°, 0°–20° S, 20° S–40° S, 40° S–60° S (equal degrees);  
60° N–35.3° N, 35.3° N–16.8° N, 16.8° N–0°, 0°–16.8° S, 16.8° S–35.3° S, 35.3° S–60° S (equal areas);  
60° N–40° N, 40° N–15° N, 15° N–0°, 0°–15° S, 15° S–40° S, 40° S–60° S;  
60° N–35° N, 35° N–20° N, 20° N–0°, 0°–20° S, 20° S–35° S, 35° S–60° S;  
60° N–30° N, 30° N–15° N, 15° N–0°, 0°–15° S, 15° S–30° S, 30° S–60° S.

The SST proxy records have limited geographical distribution, variable length and variable resolution. In particular, the records are clustered in space and have a non-random spatial distribution (Extended Data Figs 1a, 2). Thus, I explore two approaches in analysing the records. In the first, I include all 61 SST proxy records and weight them equally. In the second approach, I identify locations where there is more than one record within a circle of radius 5° latitude/longitude. I identify 11 such clusters that include 43 proxy records in total. For each cluster, I estimate the mean value over time and the variation across the cores. I plot the resulting 29 records (11 clusters and 18 independent proxy records) for this 'clustered' analysis in Extended Data Fig. 1b.

Because this research is focused on change in SST not absolute values of SST, it is important that the SST reconstructions are normalized to change from present before they are summed within a latitudinal zone. The rest of this research defines present as the mean value 0–5 ka. However, in this particular instance, 8 of the SST reconstructions have their first estimate between 5 and 8 ka. Rather than assume an estimate for 5 ka, this research uses the mean value 0–10 ka when estimating the latitudinal zone averages. Once the weighted average of the zones is estimated, the deviation from present is then recalculated to be the mean value 0–5 ka to ensure consistency.

This analysis uses Monte Carlo-style simulations to estimate several sources of uncertainty from the proxy reconstruction by adding random error to each reconstruction from the estimated proxy uncertainty and the estimated uncertainty introduced from absolute dating (see previous discussion for those estimates) in each simulation. I explore structural uncertainty in the averaging method by randomly resampling the different latitudinal zonal boundaries. I also randomly resample the proxy cores to explore the uncertainty introduced from a particular set of proxy records. The three general sources of uncertainty—from the proxy reconstructions, from the latitudinal zone boundaries, and from the sample of records—are all major contributors to the final uncertainty distribution. Random resampling of the SST proxy records is the largest contributor to the uncertainty. In total, I calculate 4,000 simulations for time series of average SST over 60° N–60° S. I repeat this approach with the 29 records from the clustering procedure described above, and find similar results for the final GAST estimate (Extended Data Fig. 3a, b).

**Scaling from latitudinal SST averages to GAST.** As discussed previously, this analysis uses the PMIP global climate models<sup>31,32</sup> to scale regional SST averages to GAST. This is necessary because there are insufficient land temperature proxies available over the past 2 Myr. I use model simulations of the LGM because they most closely compare to the large changes seen in the reconstructions. For a specific LGM climate model simulation, I estimate the change in temperature between the LGM and pre-industrial runs for both SST and surface air temperature. I then estimate the ratio of change in GAST to change in 60° N–60° S average SST to be used as a scaling factor (Extended Data Fig. 4). The analyses are performed for all available model simulations using the PMIP2 database from 30 July 2009 (<https://pmip2.lscce.fr/database/access/request.shtml>); and the PMIP3 database from CMIP5 archives at PCMDI from 10 December 2015 ([http://cmip-pcmdi.llnl.gov/cmip5/data\\_portal.html](http://cmip-pcmdi.llnl.gov/cmip5/data_portal.html)). Because of the uncertainty in applying these scalars to estimate GAST, I more than double the standard deviation of the estimates from the sample of nine models (0.14) and use a scaling factor of 1.9 (1.5–2.3, 95% interval), assuming a normal distribution that includes 8 of the 9 models in the middle 67%. The uncertainty in the scaling factor causes approximately a doubling of the standard deviation of the final GAST reconstruction at each point in time.

This research uses 60° N–60° S because that is the maximum extent of the SST reconstructions. However, the high latitude Southern Hemisphere reconstructions are fairly short, and for most of the past 2 Myr, the highest latitude SST reconstruction is at 43° S in contrast to 58° N in the Northern Hemisphere. Sensitivity analysis finds that repeating the analysis with the PMIP models using 50° N–50° S finds a very similar range of scalar factors (the mean value changes by only 1%). Therefore, this research continues to use the scaling factor estimate of 1.9 (1.5–2.3, 95% interval) as equally applying to a range of 60° N–60° S or 50° N–50° S.

To estimate GAST reconstructions, I use Monte Carlo-style simulations to propagate all the previously mentioned sources of uncertainty. I sample from the simulations of 60° N–60° S average SST described previously and apply a scaling factor randomly chosen from the uncertainty distribution to calculate 5,000 simulations for the potential time series of GAST (Fig. 1a). The final GAST simulation ensemble of potential time series includes propagation from simulating the potential errors in each proxy reconstruction, from resampling the proxy records, and from randomly varying the spatial weighting methods and the scaling factors of regional average SST to GAST. When simulating GAST time series, the time series stops when any latitudinal zone no longer has any reconstructions available. Therefore, the lengths of the individual time series within the final simulation GAST ensemble of potential time series (Extended Data Fig. 2c) are reflective of the availability of SST proxy reconstructions (Extended Data Fig. 2b).

The assumption of linear and constant scaling over the past 2 Myr is a weakness in this approach, but it is necessary because of the limited available data<sup>7,15</sup>. One way to explore how that scalar value could potentially change over time is by investigating the PMIP simulations of the mid-Pliocene warm period (mPWP, 3.0–3.3 Ma)<sup>44</sup>. An analysis of cross-model means finds a scalar of 1.6, which is 15% lower than the median estimate from the LGM PMIP experiments and is within the proposed confidence interval. Moreover, the mPWP was much warmer (1.8–3.6°C warmer than present) with much less ice and sea levels higher by  $22 \pm 10$  m (ref. 44). At the Pleistocene transition at ~2.7 Ma, there was a substantial increase in Northern Hemisphere ice sheets, and by 2.4 Ma, the climate transitioned to ~41-kyr glacial cycles until the MPT<sup>45</sup>. Therefore, the Earth's climate over the past 2 Myr was much more similar to the LGM than to the mPWP. Thus, if the scalar is state dependent, as the model simulations suggest, the value at 2 Ma is likely to be closer to 1.9 than 1.6 and thus well within the proposed interval.

To test the impact of the assumption of a constant linear scalar of the ratio of change in GAST to change in average SST over the latitudinal zones 60° N to 60° S, I analyse two alternative methods. Rather than use the single value of 1.9 estimated

from the LGM, I use a moving scalar that is defined by the time points of 1.9 at the LGM and 1.6 at the mPWP. Because the scalar is thought to potentially vary with climate state, I use two different time series to construct the scalar time series: reconstructions of deep sea oxygen isotopes<sup>30</sup> and of relative sea level<sup>5</sup>. I linearly scale each of these time series such that their mean value at the LGM (mean over 19–23 ka) is 1.9 and their mean value at the mPWP (mean over 3–3.3 Ma) is 1.6. The resulting estimated median GAST time series are very similar to the GAST reconstruction estimated from a constant scalar: 0.998 correlation for the deep sea oxygen isotopes method and 0.998 for the relative sea level method. Investigations of polar amplification are also similar: 0.59 (0.45 to 0.79, 95% interval) for the deep sea oxygen isotopes method and 0.59 (0.45 to 0.74, 95% interval) for the relative sea level method, as compared to 0.61 (0.43 to 0.85, 95% interval) for the primary GAST reconstruction. The regressions also find that a quadratic relationship is not significant. The uncertainty estimates of the two alternative approaches are underestimates because they do not include uncertainty in the base reconstructions themselves nor do they include any uncertainty in the LGM and mPWP estimates used to scale the base reconstructions.

**Validity testing the GAST reconstruction using particular points in time.** To test the validity of the GAST reconstruction, the new record can be compared to previous published reconstructions for points of interest, such as the LGM or the Last Interglacial. The new GAST reconstruction finds global cooling at the LGM (~21 ka) of 6.2°C (4.5–8.1°C, 95% interval) from the present value. This estimate is similar to the recent IPCC synthesis of the 'very likely' range (>95% probability) of 3–8°C (ref. 1). The present analysis finds a higher most-likely value consistent with other recent proxy-based reconstructions of the LGM<sup>17,46,47</sup>. Lower model-based estimates of LGM cooling could be influenced by underestimates of the changes in LGM radiative forcing, such as from changes in dust and vegetation<sup>1,48</sup>, or underestimates of climate sensitivity at the LGM. The GAST estimate for maximum warming during the Last Interglacial (~125 ka) is 2.0°C (0.4–3.6°C, 95% interval) warmer than present. This result is consistent with recent proxy estimates: sea level rise is likely to have exceeded 8 m above present levels<sup>49</sup>, a large warming over Antarctica during the Last Interglacial<sup>50</sup>, and a GAST estimate of 1.9°C warming above pre-industrial levels<sup>51</sup>. The present estimate for maximum warming during the Last Interglacial is higher than some model simulations<sup>52</sup>. Comparisons of the new GAST reconstruction with additional palaeoclimatic reconstructions find strong correlations, as would be expected (Extended Data Table 3).

**Validity testing the GAST reconstruction using PMIP model outputs.** The PMIP model outputs are used to test the robustness of this research's method for estimating GAST. The question explored is: if the proposed approach of this paper is applied to grid cell values from the models for the same locations as the SST reconstructions, how would the estimated GAST value compare to the model's GAST value? For each of the nine PMIP model simulations available, nearest neighbour classification (via the *knn1* function in the *class* package<sup>53</sup> in the R statistical program, <http://cran.r-project.org/web/packages/class/index.html>) is used to identify the grid cell closest to each SST proxy reconstruction. The change in SST from present to the LGM at those locations is used as an input to the GAST estimation methods described above. The models' surface air temperature outputs are used to directly estimate GAST for each model. The final estimates of GAST using the 61 SST proxy reconstruction locations are then compared to the models' GAST values (Extended Data Fig. 5). The same procedure is repeated with only the locations of the 5 SST reconstructions that cover the full past 2 Myr (Extended Data Fig. 5). The median estimates for change in GAST at the LGM using PMIP model outputs are very similar (–4.5°C directly from the models, versus –5.1°C from the 61 record locations and –4.9°C from the 5 record locations) when combined across the 9 models, and well within the estimated 95% interval. The method of this paper does find a larger range of uncertainty in GAST, as to be expected when comparing 61 or 5 grid cells to the full surface air temperature model outputs (Extended Data Fig. 5). Based on these 9 models that are available, the median estimates suggest the new GAST reconstruction may overestimate cooling by 10%, but the median estimates vary in direction and magnitude across the individual models and thus such a conclusion is specific to the set of model outputs.

**Robustness of the GAST reconstruction to particular sample of cores.** Early in the reconstruction the estimate is based on 61 reconstructions, but at 400 ka on only 14 reconstructions, at 800 ka on only 8 reconstructions, and at 2 Ma on only 5 reconstructions (Extended Data Fig. 2). This research uses three different analyses to assess the robustness of the final GAST reconstruction to the particular set of SST reconstructions currently available. First, the final GAST ensemble includes bootstrap Monte Carlo-style simulations that resample the reconstructions before following the methods to calculate GAST. The final ensemble of GAST time series are directly from this bootstrap simulation, and the overall uncertainty thus includes the uncertainty caused by the particular set of 61 reconstructions. Second, the entire methodology, including the bootstrap simulations, is repeated



for just the 11 clusters plus 18 individual records, as described above. The entire methodology, including the bootstrap simulations, is also repeated for just the 5 SST reconstructions that cover the full past 2 Myr. Extended Data Fig. 3a, b compares the median estimates from those two variations to the primary GAST reconstruction. Although the reduction in number of reconstructions causes larger uncertainty and potential different variance structures, the median estimates are very similar (0.997 correlation for the clustered version and 0.953 correlation for the 5-record version). Third, as described previously, this analysis's methodology is applied to the PMIP model outputs for both the full 61 reconstructions and just the 5 records that cover the full past 2 Myr and finds GAST estimates consistent with the models' air surface temperature outputs (Extended Data Fig. 5).

**DWT reconstructions.** This analysis compares GAST to 12 different proxy-based reconstructions of changes in DWT developed from three different methods (Extended Data Table 2). Elderfield *et al.*<sup>4</sup> estimate DWT using Mg/Ca ratios from bottom-dwelling foraminifera. Rohling *et al.*<sup>5</sup> estimate sea level using surface planktonic oxygen isotopes from Mediterranean Sea sediments and then use the sea level estimate to remove the ice-volume effect from the global benthic oxygen stable isotope data to estimate global DWT changes. Bates *et al.*<sup>6</sup> estimate DWT from oxygen isotope records from benthic foraminifera shells using regression analysis for ten different deep-ocean records, using a much simpler model for the relationship between benthic oxygen isotopes and DWT as well as older age models than the other two reconstructions.

**GHG radiative forcing.** For GHG concentrations from the past 800 kyr, I use stacked reconstructions from Antarctic ice cores<sup>14,18,54</sup>. The same 1-kyr interpolation method is applied to these records as described above for the SST reconstructions. I calculate the radiative forcing changes for CO<sub>2</sub> and methane from the past 800 kyr using the equations from Kohler *et al.*<sup>17</sup> and Hansen *et al.*<sup>11</sup> of total forcing =  $\alpha(\text{CO}_2 \text{ forcing} + \beta \times \text{methane forcing})$ . I apply highly conservative 95% intervals for the parameters in the proposed approximations:  $\alpha$  (approximation for N<sub>2</sub>O) is 1.12 (1.0–1.24, 95% interval),  $\beta$  (efficacy of methane) is 1.4 (1.0–1.8, 95% interval), and I apply conservative uncertainty of 20% (95% interval) for the overall equation.

**Regression analyses.** To investigate the relationship between the new GAST reconstruction and other palaeoclimatic reconstructions (for example, Extended Data Table 3 and lines in Fig. 2), I estimate 'GAST sensitivity' (the estimated linear relationship of change in GAST for each unit of change in the palaeoclimate record) and the correlation between the two reconstructions. I first randomly sample a single time series from the GAST simulation ensemble of potential time series (described above) and a single time series from the simulation ensemble for the comparison record. I then normalize each record to be deviations from present, where present is defined as the mean value over 1–5 ka. It is necessary to use 5 ka because not all reconstructions have estimates for 1–3 ka. I use weighted least squares regressions without an intercept, because both records are deviations from present. I test both linear functions and nonlinear, such as quadratic, relationships and I use an ANOVA test to assess the improved fit of alternative functional relationships. I also quantify the correlation of the two time series. I then repeat the analyses for at least 500 random draws of time series from each reconstruction. Because there is high autocorrelation in most palaeoclimatic reconstructions, I include an autoregressive model to evaluate the potential underestimation of error in the regression coefficients that can be caused by autocorrelation<sup>55</sup>. For example, the autoregressive model for GAST as a function of radiative forcing from GHGs does not significantly change the regression coefficient, as predicted by theory, but it does increase the estimated standard error of the regression coefficient from 1.8% to 3.5% of the median value. However, the regression standard error is an insignificant contribution to the overall uncertainty analysis, which estimates an analogous standard error of 19% overall. To assess whether the regression results vary for deep glacial states, I define deep glacial periods as the bottom 25% of the comparison record (less than  $-2 \text{ W m}^{-2}$  for GHG radiative forcing) and repeat the analyses with the two sets of time periods separately. I convert the estimate of change in GAST to change in GHG radiative forcing to ESS by multiplying by  $3.7 \text{ W m}^{-2}$  (the change in radiative forcing from a doubling of CO<sub>2</sub>)<sup>2</sup>.

Investigations of ESS are limited by the availability of GHG reconstructions. Because reconstructions of atmospheric CO<sub>2</sub> before 800 ka are highly uncertain and limited in temporal resolution and reconstructions of methane before 800 ka do not exist, this research focuses on analysis of the past 800 kyr of GHG reconstructions. However, I analyse the reconstructions of CO<sub>2</sub> based on boron isotopes<sup>56</sup> that are available for limited time points across the past 2 Myr. I use the equation from Kohler *et al.*<sup>17</sup> to estimate CO<sub>2</sub> radiative forcing. I repeat the regression analyses described above for the limited set of time points to estimate GAST sensitivity to CO<sub>2</sub> radiative forcing and I repeat the analysis separately for 0–1 Ma and 1–2 Ma (Extended Data Table 3 and Extended Data Fig. 7). For the past 1 Myr, the results are substantially similar to the results obtained from a

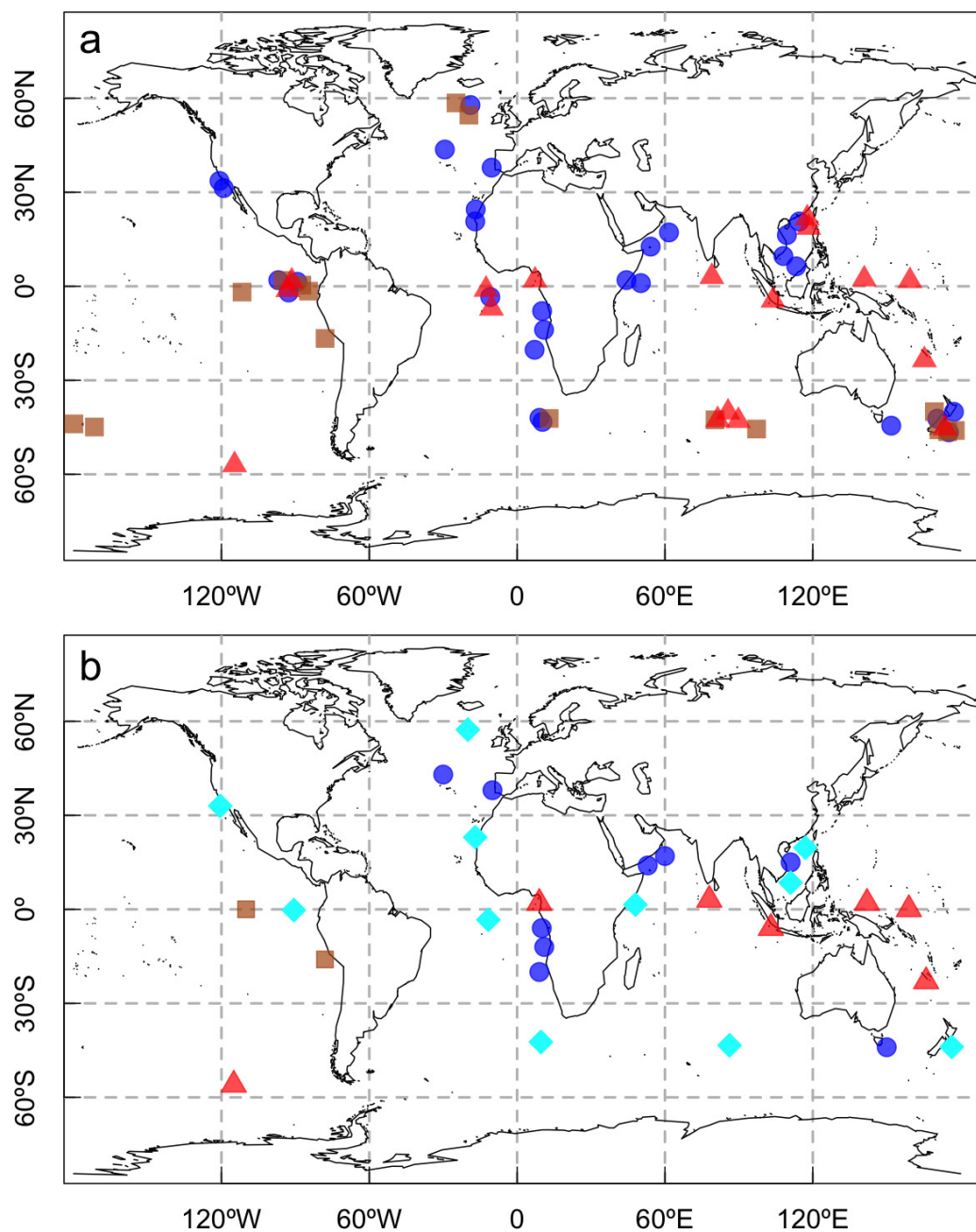
comparison with CO<sub>2</sub> reconstructed from ice cores, with lower correlation and coefficient estimates as would be expected due to the higher uncertainty in the CO<sub>2</sub> reconstruction. Prior to 1 Ma, CO<sub>2</sub> radiative forcing from boron isotopes is poorly correlated with GAST changes, suggesting either a decoupling of GAST and CO<sub>2</sub> before the MPT, or more likely, errors in the CO<sub>2</sub> reconstruction, GAST reconstruction, and/or the relative dating of the records.

**Probabilistic breakpoint identification.** This research uses probabilistic breakpoint identification to estimate changes in GAST time trends over the past 2 Myr. Breakpoint simulation detects and identifies changes within time series by decomposing the time series into linear trends and breakpoints. This research uses the *bfast* function from the *bfast* package<sup>57</sup> in the R statistical program (<http://cran.r-project.org/web/packages/bfast/index.html>), which iteratively estimates time trends and break points through a piecewise linear trend to identify optimal values. The *bfast* simulation program is applied independently to 500 randomly selected time series from the final ensemble of GAST time series to estimate the empirically-fitted frequency distributions in Fig. 3. The program also estimates the time trend before and after the breakpoint. No smoothing is used in this analysis.

**Data and code availability.** Supplementary Methods includes R code for key methods described in the paper. Supplementary Data includes the new GAST reconstruction at 2.5%, 5%, 25%, 50%, 75%, 95% and 97.5% likelihood values, and the 61 SST reconstructions used to create the GAST reconstruction, including a detailed summary table.

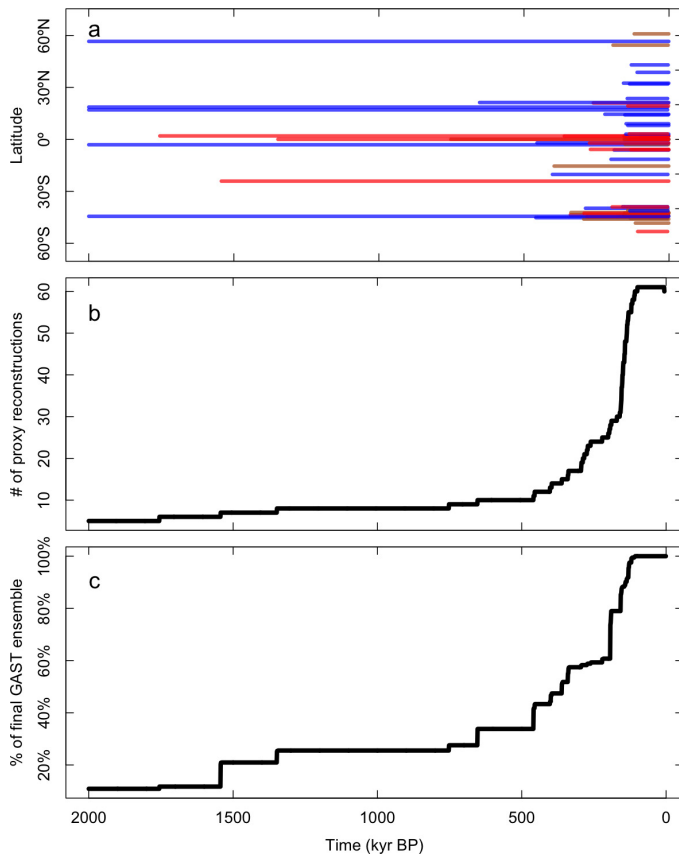
1. Braconnot, P. *et al.* Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum — Part 1: experiments and large-scale features. *Clim. Past* **3**, 261–277 (2007).
2. Harrison, S. P. *et al.* Climate model benchmarking with glacial and mid-Holocene climates. *Clim. Dyn.* **43**, 671–688 (2014).
3. Hargreaves, J. C., Annan, J. D., Yoshimori, M. & Abe-Ouchi, A. Can the Last Glacial Maximum constrain climate sensitivity? *Geophys. Res. Lett.* **39**, L24702 (2012).
4. Mix, A. C., Bard, E. & Schneider, R. Environmental processes of the ice age: land, oceans, glaciers (EPILOG). *Quat. Sci. Rev.* **20**, 627–657 (2001).
5. Müller, P. J., Kirst, G., Ruhland, G., Von Storch, I. & Rossel-Mele, A. Calibration of the alkenone paleotemperature index U<sub>37</sub><sup>K</sup> based on core-tops from the eastern South Atlantic and the global ocean (60°N–60°S). *Geochim. Cosmochim. Acta* **62**, 1757–1772 (1998).
6. Herbert, T. D. Review of alkenone calibrations (culture, water column, and sediments). *Geochem. Geophys. Geosyst.* **2**, <http://dx.doi.org/10.1029/2000GC000055> (2001).
7. Mashiotta, T. A., Lea, D. W. & Spero, H. J. Glacial-interglacial changes in Subantarctic sea surface temperature and  $\delta^{18}\text{O}$ -water using foraminiferal Mg. *Earth Planet. Sci. Lett.* **170**, 417–432 (1999).
8. Elderfield, H. & Ganssen, G. Past temperature and  $\delta^{18}\text{O}$  of surface ocean waters inferred from foraminiferal Mg/Ca ratios. *Nature* **405**, 442–445 (2000).
9. Barrows, T. T., Juggins, S., De Deckker, P., Calvo, E. & Pelejero, C. Long-term sea surface temperature and climate change in the Australian-New Zealand region. *Paleoceanography* **22**, PA2215 (2007).
10. Haam, E. & Huybers, P. A test for the presence of covariance between time-uncertain series of data with application to the Dongge Cave speleothem and atmospheric radiocarbon records. *Paleoceanography* **25**, PA001713 (2010).
11. Lin, L., Khider, D., Lisiecki, L. E. & Lawrence, C. E. Probabilistic sequence alignment of stratigraphic records. *Paleoceanography* **29**, 976–989 (2014).
12. Martinson, D. G. *et al.* Age dating and the orbital theory of the ice ages: development of a high-resolution 0 to 300,000-year chronostratigraphy. *Quat. Res.* **27**, 1–29 (1987).
13. Huybers, P. Glacial variability over the last two million years: an extended depth-derived age model, continuous obliquity pacing, and the Pleistocene progression. *Quat. Sci. Rev.* **26**, 37–55 (2007).
14. Haywood, A. M. Large-scale features of Pliocene climate: results from the Pliocene Model Intercomparison Project. *Clim. Past* **9**, 191–209 (2013).
15. Bell, D. B., Jung, S. J. A. & Kroon, D. The Plio-Pleistocene development of Atlantic deep-water circulation and its influence on climate trends. *Quat. Sci. Rev.* **123**, 265–282 (2015).
16. Kageyama, M. *et al.* The Last Glacial Maximum climate over Europe and western Siberia: a PMIP comparison between models and data. *Clim. Dyn.* **17**, 23–43 (2001).
17. Ballantyne, A. P., Lavine, M., Crowley, T. J., Liu, J. & Baker, P. B. Meta-analysis of tropical surface temperatures during the Last Glacial Maximum. *Geophys. Res. Lett.* **32**, L05712 (2005).
18. Schneider von Deimling, T., Ganopolski, A., Held, H. & Rahmstorf, S. How cold was the Last Glacial Maximum? *Geophys. Res. Lett.* **33**, L14709 (2006).
19. Kopp, R. E., Simons, F. J., Mitrovica, J. X., Maloof, A. C. & Oppenheimer, M. Probabilistic assessment of sea level during the last interglacial stage. *Nature* **462**, 863–867 (2009).
20. Sime, L. C., Wolff, E. W., Oliver, K. I. C. & Tindall, J. C. Evidence for warmer interglacials in East Antarctic ice cores. *Nature* **462**, 342–345 (2009).
21. Turney, C. S. & Jones, R. T. Does the Agulhas Current amplify global temperatures during super-interglacials? *J. Quat. Sci.* **25**, 839–843 (2010).

52. Otto-Bliesner, B. L. *et al.* How warm was the last interglacial? New model–data comparisons. *Phil. Trans. R. Soc.* **371**, <http://dx.doi.org/10.1098/rsta.2013.0097> (2013).
53. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* 4th edn (Springer, 2002).
54. Loulergue, L. *et al.* Orbital and millennial-scale features of atmospheric CH<sub>4</sub> over the past 800,000 years. *Nature* **453**, 383–386 (2008).
55. Abraham, B. & Ledolter, J. *Introduction to Regression Modeling* (Duxbury Press, 2006).
56. Honisch, B., Hemming, G., Archer, D., Siddall, M. & McManus, J. Atmospheric carbon dioxide concentration across the Mid-Pleistocene Transition. *Science* **324**, 1551–1554 (2009).
57. Verbesselt, J., Hyndman, R., Newnham, G. & Culvenor, D. Detecting trend and seasonal changes in satellite image time series. *Remote Sens. Environ.* **114**, 106–115 (2010).
58. de Garidel-Thoron, T., Rosenthal, Y., Bassinot, F. & Beaufort, L. Stable sea surface temperatures in the western Pacific warm pool over the past 1.75 million years. *Nature* **433**, 294–298 (2005).
59. Lea, D. W. The 100 000-yr cycle in tropical SST, greenhouse forcing, and climate sensitivity. *J. Clim.* **17**, 2170–2179 (2004).
60. Lea, D. W. *et al.* Paleoclimate history of Galapagos surface waters over the last 135,000 yr. *Quat. Sci. Rev.* **25**, 1152–1167 (2006).
61. Medina-Elizalde, M. & Lea, D. W. The mid-Pleistocene transition in the tropical Pacific. *Science* **310**, 1009–1012 (2005).
62. Mohtadi, M. *et al.* Late Pleistocene surface and thermocline conditions of the eastern tropical Indian Ocean. *Quat. Sci. Rev.* **29**, 887–896 (2010).
63. Nürnberg, D., Müller, A. & Schneider, R. R. Paleo-sea surface temperature calculations in the equatorial east Atlantic from Mg/Ca ratios in planktic foraminifera: A comparison to sea surface temperature estimates from U<sub>37</sub><sup>K</sup>, oxygen isotopes, and foraminiferal transfer function. *Paleoceanography* **15**, 124–134 (2000).
64. Oppo, D. W. & Sun, Y. B. Amplitude and timing of sea-surface temperature change in the northern South China Sea: dynamic link to the East Asian monsoon. *Geology* **33**, 785–788 (2005).
65. Pahnke, K., Zahn, R., Elderfield, H. & Schulz, M. 340,000-year centennial-scale marine record of southern hemisphere climatic oscillation. *Science* **301**, 948–952 (2003).
66. Rickaby, R. E. M. & Elderfield, H. Planktonic foraminiferal Cd/Ca: paleonutrients or paleotemperature? *Paleoceanography* **14**, 293–303 (1999).
67. Russon, T. *et al.* Inter-hemispheric asymmetry in the early Pleistocene Pacific warm pool. *Geophys. Res. Lett.* **37**, L11601 (2010).
68. Saraswat, R., Nigam, R., Weldeab, S., Mackensen, A. & Naidu, P. D. A first look at past sea surface temperatures in the equatorial Indian Ocean from Mg/Ca in foraminifera. *Geophys. Res. Lett.* **32**, L24605 (2005).
69. Wei, G. J., Deng, W. F., Liu, Y. & Li, X. H. High-resolution sea surface temperature records derived from foraminiferal Mg/Ca ratios during the last 260 ka in the northern South China Sea. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **250**, 126–138 (2007).
70. Weldeab, S., Lea, D. W., Schneider, R. R. & Andersen, N. 155,000 years of West African monsoon and ocean thermal evolution. *Science* **316**, 1303–1307 (2007).
71. Brathauer, U. & Abelmann, A. Late Quaternary variations in sea surface temperatures and their relationship to orbital forcing recorded in the Southern Ocean (Atlantic sector). *Paleoceanography* **14**, 135–148 (1999).
72. Kandiano, E. S., Bauch, H. A. & Müller, A. Sea surface temperature variability in the North Atlantic during the last two glacial-interglacial cycles: comparison of faunal, oxygen isotopic, and Mg/Ca-derived records. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **204**, 145–164 (2004).
73. Labeyrie, L. *et al.* Hydrographic changes of the Southern Ocean (southeast Indian sector) over the last 230 kyr. *Paleoceanography* **11**, 57–76 (1996).
74. Pisias, N. G. & Mix, A. C. Spatial and temporal oceanographic variability of the eastern equatorial Pacific during the late Pleistocene: evidence from Radiolaria microfossils. *Paleoceanography* **12**, 381–393 (1997).
75. Weaver, P. P. E., Carter, L. & Neil, H. L. Response of surface water masses and circulation to Late Quaternary climate change east of New Zealand. *Paleoceanography* **13**, 70–83 (1998).
76. Weaver, P. P. E. *et al.* Combined coccolith, foraminiferal, and biomarker reconstruction of paleoceanographic conditions over the past 120 kyr in the northern North Atlantic (59°N, 23°W). *Paleoceanography* **14**, 336–349 (1999).
77. Bard, E. Climate shock — Abrupt changes over millennial time scales. *Phys. Today* **55**, 32–38 (2002).
78. Bard, E., Rostek, F. & Sonzogni, C. Interhemispheric synchrony of the last deglaciation inferred from alkenone palaeothermometry. *Nature* **385**, 707–710 (1997).
79. Clemens, S. C., Prell, W. L., Sun, Y., Liu, Z. & Chen, G. Southern Hemisphere forcing of Pliocene δ<sup>18</sup>O and the evolution of Indo-Asian monsoons. *Paleoceanography* **23**, PA4210 (2008).
80. Herbert, T. D., Peterson, L. C., Lawrence, K. T. & Liu, Z. Tropical ocean temperatures over the past 3.5 million years. *Science* **328**, 1530–1534 (2010).
81. Dubois, N. *et al.* Millennial-scale variations in hydrography and biogeochemistry in the Eastern Equatorial Pacific over the last 100 kyr. *Quat. Sci. Rev.* **30**, 210–223 (2011).
82. Eglinton, G. *et al.* Molecular record of secular sea surface temperature changes on 100-year timescales for glacial terminations I, II and IV. *Nature* **356**, 423–426 (1992).
83. Horikawa, K., Minagawa, M., Murayama, M., Kato, Y. & Asahi, H. Spatial and temporal sea-surface temperatures in the eastern equatorial Pacific over the past 150 kyr. *Geophys. Res. Lett.* **33**, L13605 (2006).
84. Lawrence, K. T., Liu, Z. H. & Herbert, T. D. Evolution of the eastern tropical Pacific through Plio-Pleistocene glaciation. *Science* **312**, 79–83 (2006).
85. Lawrence, K. T., Herbert, T. D., Brown, C. M., Raymo, M. E. & Haywood, A. M. High-amplitude variations in North Atlantic sea surface temperature during the early Pliocene warm period. *Paleoceanography* **24**, PA2218 (2009).
86. Müller, P. J., Cepek, M., Ruhland, G. & Schneider, R. R. Alkenone and coccolithophorid species changes in late Quaternary sediments from the Walvis Ridge: implications for the alkenone paleotemperature method. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **135**, 71–96 (1997).
87. Pahnke, K. & Sachs, J. P. Sea surface temperatures of southern midlatitudes 0–160 kyr BP. *Paleoceanography* **21**, PA2003 (2006).
88. Pelejero, C., Grimalt, J. O., Heilig, S., Kienast, M. & Wang, L. J. High-resolution U<sub>37</sub><sup>K</sup> temperature reconstructions in the South China Sea over the past 220 kyr. *Paleoceanography* **14**, 224–231 (1999).
89. Pelejero, C., Calvo, E., Barrows, T. T., Logan, G. A. & De Deckker, P. South Tasman Sea alkenone palaeothermometry over the last four glacial/interglacial cycles. *Mar. Geol.* **230**, 73–86 (2006).
90. Rostek, F., Bard, E., Beaufort, L., Sonzogni, C. & Ganssen, G. Sea surface temperature and productivity records for the past 240 kyr in the Arabian Sea. *Deep-Sea Res.* **44**, 1461–1480 (1997).
91. Sachs, J. P. & Anderson, R. F. Fidelity of alkenone paleotemperatures in southern Cape Basin sediment drifts. *Paleoceanography* **18**, 1082 (2003).
92. Schneider, R. R., Müller, P. J. & Ruhland, G. Late Quaternary surface circulation in the east equatorial South Atlantic: evidence from alkenone sea surface temperatures. *Paleoceanography* **10**, 197–219 (1995).
93. Sicre, M. A. *et al.* Biomarker stratigraphic records over the last 150 kyears off the NW African coast at 25°N. *Org. Geochem.* **31**, 577–588 (2000).
94. Villanueva, J., Grimalt, J. O., Cortijo, E., Vidal, L. & Labeyrie, L. Assessment of sea surface temperature variations in the central North Atlantic using the alkenone unsaturation index U<sub>37</sub><sup>K</sup>. *Geochim. Cosmochim. Acta* **62**, 2421–2427 (1998).
95. Yamamoto, M., Oba, T., Shimamura, J. & Ueshima, T. Orbital-scale anti-phase variation of sea surface temperature in mid-latitude North Pacific margins during the last 145,000 years. *Geophys. Res. Lett.* **31**, L16311 (2004).
96. Yamamoto, M., Yamamoto, M. & Tanaka, Y. The California current system during the last 136,000 years: response of the North Pacific High to precessional forcing. *Quat. Sci. Rev.* **26**, 405–414 (2007).
97. Zhao, M. X., Huang, C. Y., Wang, C. C. & Wei, G. J. A millennial-scale U<sub>37</sub><sup>K</sup> sea-surface temperature record from the South China Sea (8°N) over the last 150 kyr: monsoon and sea-level influence. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **236**, 39–55 (2006).
98. Masson-Delmotte, V. *et al.* Atmospheric science: GRIP deuterium excess reveals rapid and orbital-scale changes in Greenland moisture origin. *Science* **309**, 118–121 (2005).
99. Masson-Delmotte, V. *et al.* Past temperature reconstructions from deep ice cores: relevance for future climate change. *Clim. Past* **2**, 145–165 (2006).

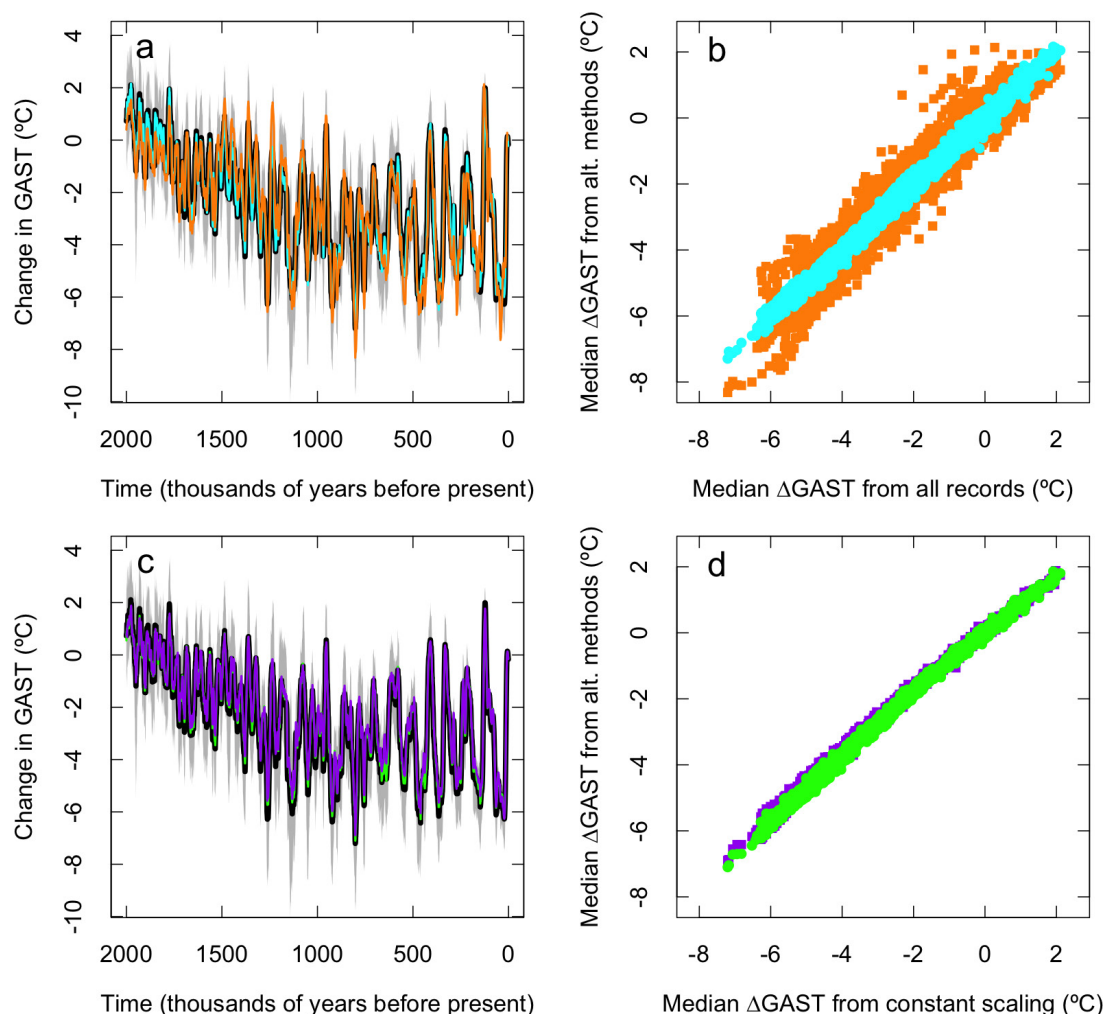


**Extended Data Figure 1 | Spatial distribution of the SST proxy reconstructions used in this analysis. a.** All 61 SST records, with methods as follows: from alkenone indices, blue circles; from Mg/Ca ratios, red triangles; and from species assemblage methods, brown squares. **b.** Repeated after clustering records within 5° latitude/longitude of each other, with the 11 clusters in cyan diamonds and the remaining 18 records as in **a**.



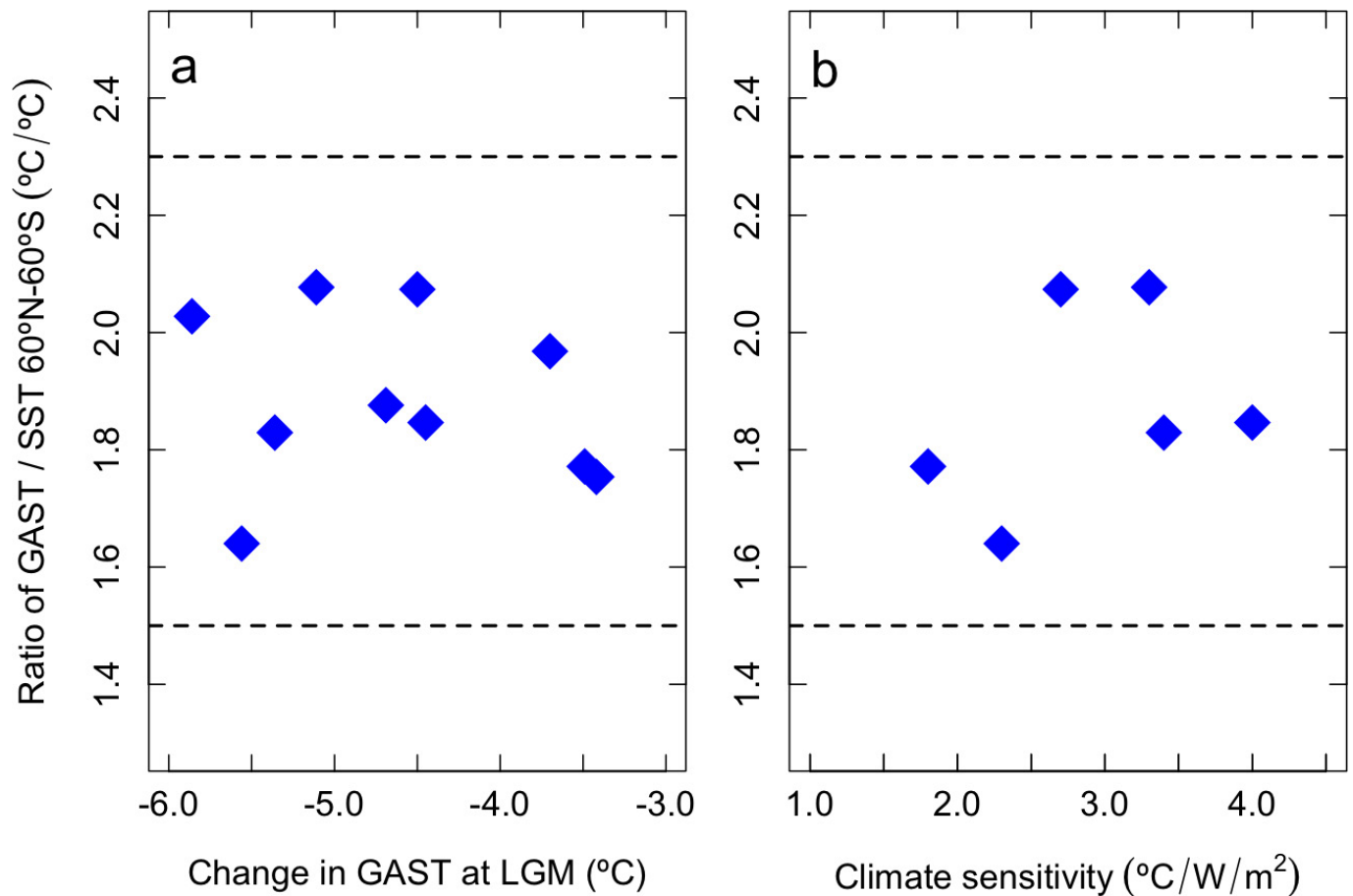


**Extended Data Figure 2 | Temporal distribution of the 61 SST proxy reconstructions used in this analysis.** **a**, Reconstruction length versus latitude, colours as in Extended Data Fig. 1. **b**, Empirical cumulative distribution function for lengths of the SST proxy reconstructions. **c**, Empirical cumulative distribution function for lengths of GAST time series in the final simulation ensemble of potential GAST time series.



**Extended Data Figure 3 | Comparison of different methods used to estimate GAST.** **a**, The primary GAST estimate (using 61 proxy reconstructions) is plotted as a function of time, with the median in black and the 95% interval in grey. The GAST estimation method is repeated for a clustering of the data (11 clusters and 18 individual reconstructions), with the median shown in cyan, and for only the 5 proxy reconstructions that cover the past 2 Myr, with the median shown in orange. **b**, The median time series from each alternative method are plotted against the primary median GAST estimate, with the clustered version in cyan circles and the 5-record version in orange squares. **c**, The primary GAST estimate

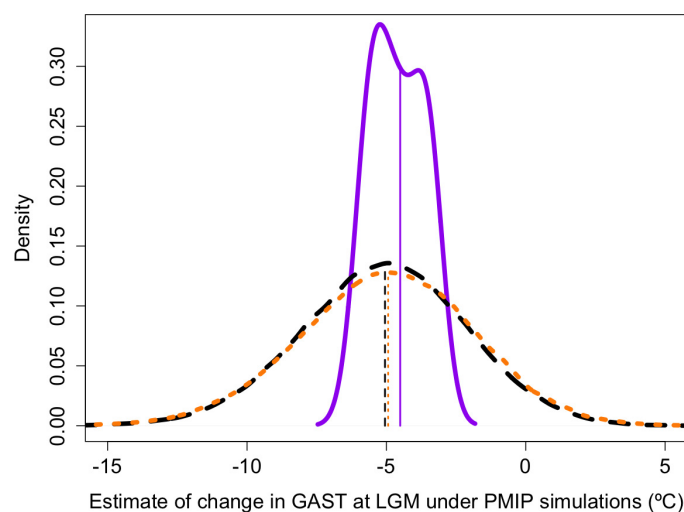
is plotted as a function of time, with the median in black and the 95% interval in grey. An alternative GAST estimation method using a time-varying scalar based on the deep-sea oxygen isotopes median estimate is shown in green, and another estimation method based on the relative sea level median estimate is shown in purple. **d**, The median time series from each alternative method is plotted against the primary median GAST estimate, with the reconstruction scaled using deep sea oxygen isotopes shown in green circles and the reconstruction scaled using relative sea level shown in purple squares.



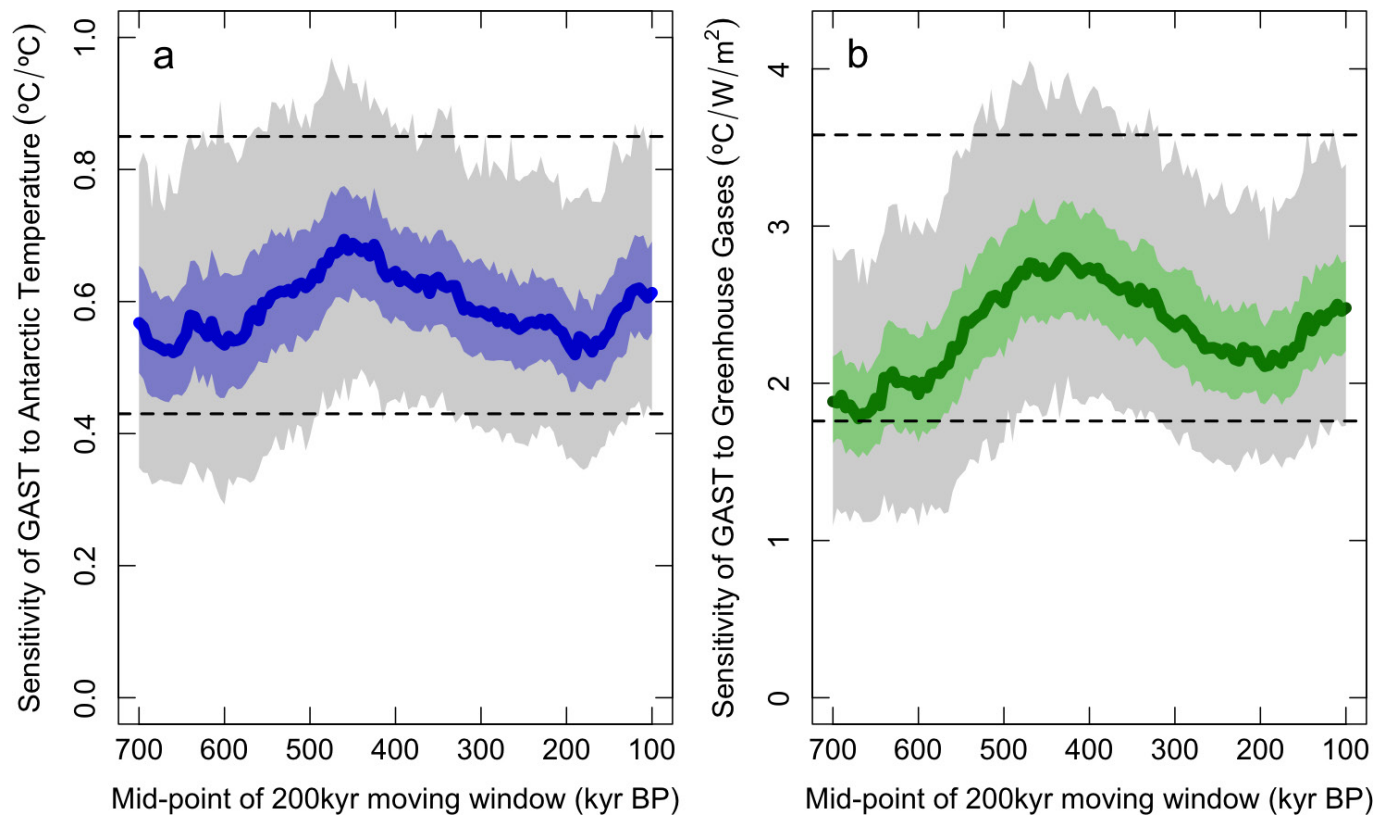
**Extended Data Figure 4 | Estimates of the ratio of change in GAST to change in average SST.** **a, b,** Scatter plots show the dependency of the ratio of change in GAST to change in average SST over the latitudinal zone 60° N to 60° S from PMIP2 and PMIP3 climate model simulations<sup>31,32</sup>

as a function of change in GAST at the LGM (**a**) and of model climate sensitivity (**b**). The climate sensitivity estimates (in °C per W m<sup>-2</sup>) are from ref. 33. Dashed lines show the scalar range used in this analysis.



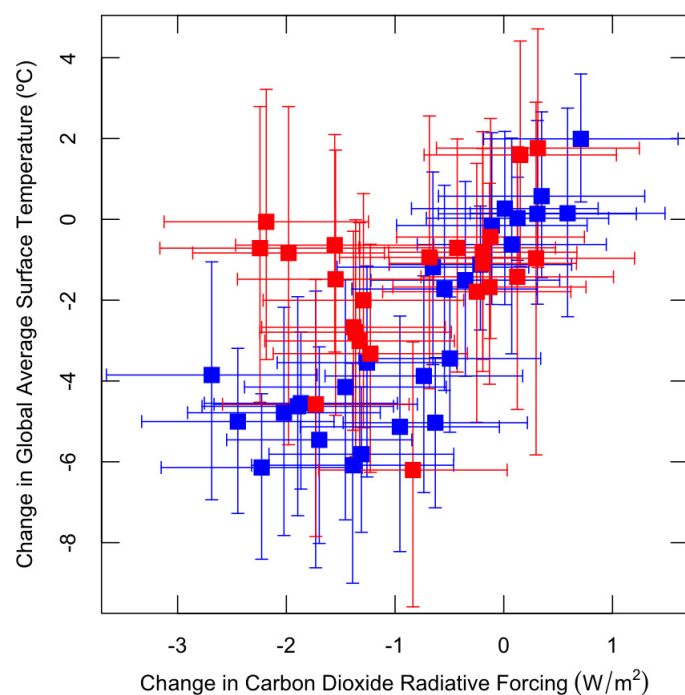


**Extended Data Figure 5 | Estimating change in GAST at the LGM using simulations drawn from PMIP model outputs.** The solid, purple line is the empirically fitted frequency distribution (shown in density on the y axis) of GAST estimated from the full air surface temperature outputs from the 9 PMIP models. The dashed, black line is the distribution of GAST estimated using the method in the present paper and the PMIP SST outputs drawn from only the locations of the 61 proxy reconstructions. The short-dashed, orange line is the same analysis completed for only the 5 proxy reconstructions that cover the past 2 Myr. The thin vertical lines are the medians of each distribution.



**Extended Data Figure 6 | The dependence of coupling relationships over time for GAST on changes in Antarctic temperature and GHG radiative forcing.** **a, b,** Regression results of change in GAST as a function of change in Antarctic temperature<sup>14</sup> (**a**) and of change in GHG radiative forcing<sup>17,18,54</sup> (**b**) are calculated for moving 200-kyr-long time windows

every 5 kyr. The solid line shows the median estimates, with the coloured and grey-shaded areas showing the 50% and 95% intervals, respectively. The dashed lines show the 95% intervals calculated from the entire time series.



**Extended Data Figure 7 | Comparison of changes in GAST to changes in  $\text{CO}_2$  radiative forcing.** Boron-isotope-based proxy reconstruction of  $\text{CO}_2$  from refs 17, 56. Blue points are from 0–1 Ma, red points are from 1–2 Ma, and error bars show 95% intervals.



**Extended Data Table 1 | Database of SST proxy reconstructions based on Mg/Ca ratio and species assemblages used in estimating GAST**

Citation	Latitude (+N/-S)	Longitude (+E/-W)	Start Date (yr BP)	End Date (yr BP)
<b><i>Mg/Ca Ratio SST Proxy Reconstructions</i></b>				
de Garidel-Thoron et al. 2005 <sup>58</sup>	2	142	7,000	1,755,000
Lea 2004 <sup>59</sup>	2	-91	1,000	361,000
Lea et al. 2006 <sup>60</sup>	0	-92	1,200	135,100
Mashiotta et al. 1999 <sup>37</sup>	-43	80	2,700	293,700
Mashiotta et al. 1999 <sup>37</sup>	-56	-115	8,110	108,450
Medina-Elizalde and Lea 2005 <sup>61</sup>	0	159	4,300	1,348,000
Mohtadi et al. 2011 <sup>62</sup>	-6	103	0	131,400
Nürnberg et al. 2000 <sup>63</sup>	-2	-12	240	274,630
Nürnberg et al. 2000 <sup>63</sup>	-6	-11	1,000	271,000
Oppo and Sun 2005 <sup>64</sup>	20	118	1,943	142,571
Pahnke et al. 2003 <sup>65</sup>	-45	175	1,951	340,835
Rickaby and Elderfield 1999 <sup>66</sup>	-40	85	4,750	196,710
Rickaby and Elderfield 1999 <sup>66</sup>	-44	90	3,600	283,670
Russon et al. 2010 <sup>67</sup>	-23	166	0	2,438,000
Saraswat et al. 2005 <sup>68</sup>	3	78	5,434	137,333
Wei et al. 2007 <sup>69</sup>	20	117	1,740	261,300
Weldeab et al. 2007 <sup>70</sup>	2	9	360	155,420
<b><i>Species Assemblages SST Proxy Reconstructions</i></b>				
Barrows et al. 2007 <sup>39</sup>	-42	170	3,830	142,712
Barrows et al. 2007 <sup>39</sup>	-46	175	5,401	133,995
Brathauer and Abelmann 1999 <sup>71</sup>	-43	12	387	338,519
Kandiano et al. 2004 <sup>72</sup>	54	-20	2,375	193,065
Labeyrie et al. 1996 <sup>73</sup> ; Rickaby and Elderfield 1999 <sup>66</sup>	-46	96	4,590	149,250
Martinson et al. 1987 <sup>42</sup>	-44	80	750	294,000
Pisias and Mix 1997 <sup>74</sup>	0	-96	360	752,047
Pisias and Mix 1997 <sup>74</sup>	-3	-83	1,914	151,051
Pisias and Mix 1997 <sup>74</sup>	0	-110	880	144,763
Pisias and Mix 1997 <sup>74</sup>	0	-86	1,479	154,653
Pisias and Mix 1997 <sup>74</sup>	-16	-78	785	466,520
Weaver et al. 1998 <sup>75</sup>	-44	-172	0	170,000
Weaver et al. 1998 <sup>75</sup>	-46	172	0	116,705
Weaver et al. 1998 <sup>75</sup>	-45	179	170	120,880
Weaver et al. 1998 <sup>75</sup>	-43	-178	0	154,600
Weaver et al. 1999 <sup>76</sup>	60	-23	0	120,000

Data are taken from refs 37, 39, 42, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75 and 76.

**Extended Data Table 2 | Database of SST proxy reconstructions based on alkenone indices used in estimating GAST**

Citation	Latitude (+N/-S)	Longitude (+E/-W)	Start Date (yr BP)	End Date (yr BP)
<i>Alkenone Indices SST Proxy Reconstructions</i>				
Bard 2002 <sup>77</sup>	38	-10	0	110,100
Bard et al. 1997 <sup>78</sup>	0	46	0	151,500
Bard et al. 1997 <sup>78</sup>	3	50	0	149,200
Clemens et al. 2008 <sup>79</sup> , Herbert et al. 2008 <sup>80</sup>	19	116	4,178	2,155,910
Dubois et al. 2011 <sup>81</sup>	2	-91	1,070	99,290
Eglinton et al. 1992 <sup>82</sup>	21	-18	736	653,042
Herbert et al. 2010 <sup>80</sup>	-2	-12	1,290	445,000
Herbert et al. 2010 <sup>80</sup>	17	60	7,110	3,330,480
Herbert et al. 2010 <sup>80</sup>	-3	-91	5,550	5,069,892
Horikawa et al. 2006 <sup>83</sup>	0	-95	1,100	154,097
Lawrence et al. 2006 <sup>84</sup>	-3	-91	5,230	5,089,802
Lawrence et al. 2009 <sup>85</sup>	58	-17	0	4,012,230
Martinez-Garcia et al. 2010 <sup>29</sup>	-43	9	0	3,642,410
Muller et al. 1997 <sup>86</sup>	-20	9	3,300	402,900
Pahnke and Sachs 2006 <sup>87</sup>	-46	175	1,950	156,360
Pahnke and Sachs 2006 <sup>87</sup>	-40	178	3,320	135,050
Pelejero et al. 1999 <sup>88</sup>	15	111	0	221,100
Pelejero et al. 1999 <sup>88</sup>	8	112	0	142,290
Pelejero et al. 2006 <sup>89</sup>	-42	170	3,570	288,540
Pelejero et al. 2006 <sup>89</sup>	-44	150	5,045	459,632
Rostek et al. 1997 <sup>90</sup>	14	53	2,300	152,600
Pahnke and Sach 2006 <sup>87</sup> , Sachs and Anderson 2003 <sup>91</sup>	-41	9	6,080	160,000
Schneider et al. 1995 <sup>92</sup>	-6	10	400	189,000
Schneider et al. 1995 <sup>92</sup>	-12	11	1,300	200,600
Sicre et al. 2000 <sup>93</sup>	25	-16	5,000	144,000
Villanueva et al. 1998 <sup>94</sup>	43	-30	2,900	285,900
Yamamoto et al. 2004, 2007 <sup>95,96</sup>	32	-119	870	157,048
Yamamoto et al. 2004, 2007 <sup>95,96</sup>	34	-122	3,186	136,475
Zhao et al. 2006 <sup>97</sup>	9	110	905	148,886

Data are taken from refs 29, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96 and 97.

Extended Data Table 3 | Comparisons of GAST with other important palaeoclimate reconstructions

Record Compared to GAST	GAST Sensitivity		Correlation	
	Median	(95% Interval)	Median	(95% Interval)
Total Greenhouse Gas Radiative Forcing <sup>14,17,18,54</sup>	2.5 °C/W/m <sup>2</sup>	(1.8, 3.6)	0.82	(0.66, 0.92)
CO <sub>2</sub> Radiative Forcing from ice cores <sup>17,18</sup>	3.0 °C/W/m <sup>2</sup>	(2.1, 4.5)	0.82	(0.63, 0.92)
CO <sub>2</sub> Radiative Forcing from Boron-isotopes, 0-1Ma <sup>17,56</sup>	2.4 °C/W/m <sup>2</sup>	(1.6, 3.4)	0.74	(0.62, 0.84)
CO <sub>2</sub> Radiative Forcing from Boron-isotopes, 1-2Ma <sup>17,56</sup>	1.3 °C/W/m <sup>2</sup>	(0.57, 2.2)	0.28	(-0.10, 0.55)
Antarctic Temperature stack <sup>14</sup>	0.61 °C/°C	(0.43, 0.85)	0.72	(0.59, 0.81)
Greenland Temperature-GRIP <sup>98</sup>	0.31 °C/°C	(0.23, 0.43)	0.89	(0.83, 0.92)
Greenland Temperature-NGRIP <sup>99</sup>	0.27 °C/°C	(0.19, 0.35)	0.91	(0.87, 0.93)
Deep-Sea Oxygen Isotope <sup>30</sup>	-4.2 °C/‰	(-5.6, -3.0)	-0.85	(-0.94, -0.58)
Deep-Water Temperature from Southwest Pacific (Core 1123) <sup>4</sup>	2.2 °C/°C	(1.5, 3.2)	0.64	(0.35, 0.78)
Deep-Water Temperature from Mediterranean Sea <sup>5</sup>	1.7 °C/°C	(0.8, 2.9)	0.47	(0.14, 0.64)
Deep-Water Temperature from Temperate North Atlantic (DSDP 607) <sup>6</sup>	1.5 °C/°C	(1.0, 2.2)	0.48	(0.36, 0.64)
Deep-Water Temperature from East Tropical Atlantic (ODP 659) <sup>6</sup>	1.5 °C/°C	(1.0, 2.1)	0.84	(0.71, 0.90)
Deep-Water Temperature from Indian Ocean (ODP 758) <sup>6</sup>	3.2 °C/°C	(2.0, 4.6)	0.27	(0.14, 0.42)
Deep-Water Temperature from Equatorial Eastern Pacific (ODP 849) <sup>6</sup>	3.0 °C/°C	(2.1, 4.2)	0.85	(0.78, 0.89)
Deep-Water Temperature from Sub-Arctic North Atlantic (ODP 980, ODP 981) <sup>6</sup>	1.4 °C/°C	(1.0, 1.9)	0.75	(0.40, 0.85)
Deep-Water Temperature from Sub-Antarctic South Atlantic (ODP 1090) <sup>6</sup>	2.1 °C/°C	(1.4, 3.0)	0.77	(0.61, 0.82)
Deep-Water Temperature from South China Sea (ODP 1143) <sup>6</sup>	2.2 °C/°C	(1.5, 3.1)	0.79	(0.64, 0.85)
Deep-Water Temperature from South China Sea (ODP 1148) <sup>6</sup>	3.5 °C/°C	(2.4, 5.0)	0.54	(0.42, 0.64)
Deep-Water Temperature from Equatorial Eastern Pacific composite (V19-30, ODP 677, ODP 846) <sup>6</sup>	2.4 °C/°C	(1.6, 3.3)	0.60	(0.37, 0.69)
Deep-Water Temperature from Southwest Pacific (Core 1123) <sup>6</sup>	1.9 °C/°C	(1.3, 2.6)	0.81	(0.58, 0.93)

'GAST sensitivity' is the estimated linear relationship of change in GAST for each unit of change in the palaeoclimate record. Data are taken from refs 4, 5, 6, 14, 17, 18, 30, 54, 56, 98 and 99.



# Progressive incision of the Channeled Scablands by outburst floods

Isaac J. Larsen<sup>1,2</sup> & Michael P. Lamb<sup>2</sup>

**The surfaces of Earth and Mars contain large bedrock canyons that were carved by catastrophic outburst floods<sup>1,2</sup>. Reconstructing the magnitude of these canyon-forming floods is essential for understanding the ways in which floods modify planetary surfaces<sup>1,2</sup>, the hydrology of early Mars<sup>3</sup> and abrupt changes in climate<sup>4</sup>. Flood discharges are often estimated by assuming that the floods filled the canyons to their brims with water; however, an alternative hypothesis is that canyon morphology adjusts during incision such that bed shear stresses exceed the threshold for erosion by a small amount<sup>5</sup>. Here we show that accounting for erosion thresholds during canyon incision results in near-constant discharges that are five- to ten-fold smaller than full-to-the-brim estimates for Moses Coulee, a canyon in the Channeled Scablands, which was carved during the Pleistocene by the catastrophic Missoula floods in eastern Washington, USA. The predicted discharges are consistent with flow-depth indicators from gravel bars within the canyon. In contrast, under the assumption that floods filled canyons to their brims, a large and monotonic increase in flood discharge is predicted as the canyon was progressively incised, which is at odds with the discharges expected for floods originating from glacial lake outbursts. These findings suggest that flood-carved landscapes in fractured rock might evolve to a threshold state for bedrock erosion, thus implying much lower flood discharges than previously thought.**

Field investigation of the Channeled Scablands, which contains vast tracts of scoured bedrock, giant gravel bars and deep canyons, led to the eventual recognition that short-lived, catastrophic floods, rather than uniformitarian processes, were the dominant agents of canyon formation in the region<sup>1,6</sup>. Outburst floods caused by the sudden drainage of lakes due to the failure of glacial or other natural dams have been documented in many settings on Earth<sup>7</sup>; and enormous bedrock canyons, morphologically similar to those in the Channeled Scablands, provide evidence for widespread outburst flooding on the surface of Mars<sup>2</sup>. Canyons carved into planetary surfaces are hence a key record of the palaeo-hydrosphere<sup>3</sup>, and constraining the magnitude of the canyon-carving floods is required to understand whether megafloods—immense, short-lived, catastrophic floods—have the potential to trigger climate change, by altering ocean circulation on Earth<sup>4</sup> or by episodically generating oceans on Mars<sup>8</sup>.

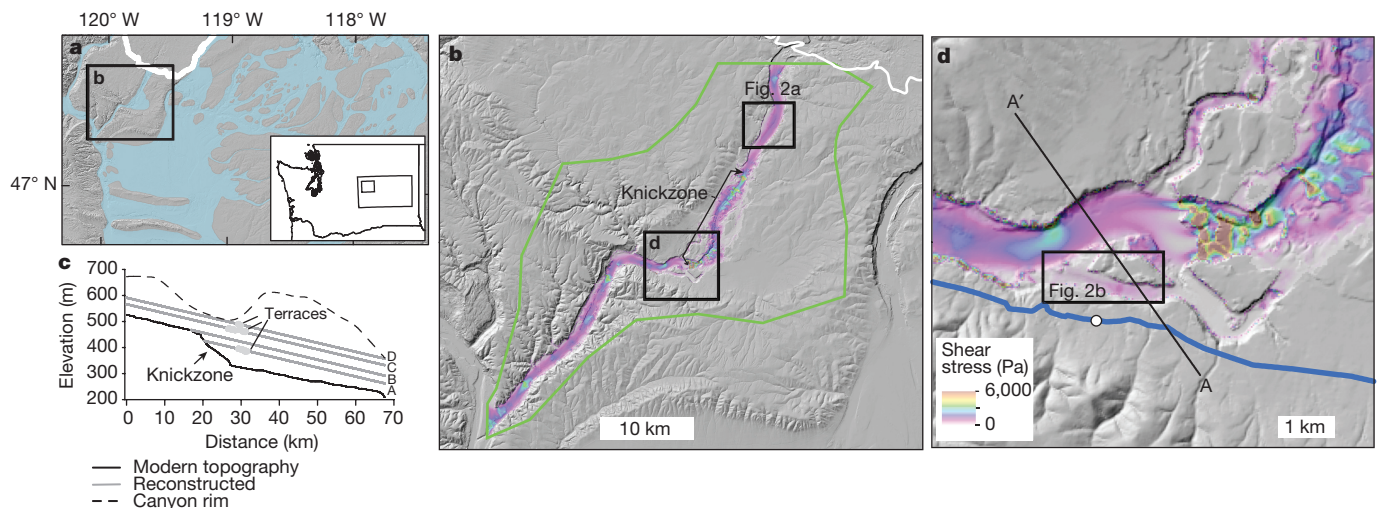
Although bedrock canyons and associated depositional landforms provide direct evidence for palaeo-floods, reconstructing the hydraulics of outburst floods is challenging. Outburst floods often initiate subglacially<sup>9</sup> or, in the case of Mars, from subsurface aquifers<sup>10</sup>; hence, canyon geometry and the grain size of flood-transported sediment often provide the only clues from which water depth, bed shear stress and discharge can be reconstructed. Outburst floods in the Channeled Scablands and on Mars have traditionally been assumed to fill the canyons up to the elevation of high-water marks such as eroded channel margins<sup>11</sup> along the rims of the canyons. Consequently, most hydraulic modelling efforts implicitly assume that the canyons were filled with water to their brims<sup>11–14</sup> or attempt to match modelled

flood depths to the brim-full markers<sup>12,14,15</sup>. However, the actual flow depth is unknown<sup>10</sup> because canyon floors must have progressively lowered as a result of bedrock incision, and high-water markers may have been active only early in canyon formation; therefore, the discharges predicted from ‘brim-full’ models are reported as maximum estimates<sup>11,13</sup>.

In the case of other bedrock canyons that formed over millions of years, such as the Grand Canyon, USA, water flows that were responsible for carving the canyon had depths that were always a small fraction of the canyon relief. Recent mechanistic studies<sup>5,9,16–20</sup> and theory for bedrock erosion by plucking of blocks from the bed<sup>21</sup> or toppling at waterfalls<sup>22,23</sup> indicate that, where rock is fractured, bedrock channels will incise via the entrainment of bedrock blocks from the bed when fluid stresses exceed the threshold for entrainment. This ‘threshold shear stress’ model implies that flows in outburst-flood-carved canyons will drop below brim-full as canyons deepen. Resolving whether canyons carved by outburst floods should be interpreted as channels filled to the brim or as valleys with flow only near their bottoms is key for reconstructing palaeo-flood discharge. However, evaluating these two ‘end-member’ models is challenging; both require removal of rock, but information on how canyons evolve during formation is lacking.

We conduct a quantitative evaluation of the end-member brim-full and threshold shear stress models by numerically simulating floods in Moses Coulee, a 70-km-long canyon in the Channeled Scablands (Fig. 1) with evidence of flooding 175 m above the canyon floor and a history of at least four late-Pleistocene-epoch outburst floods<sup>24</sup>. We selected Moses Coulee for the study because, unlike other canyons in the Channeled Scablands, the planform geometry is relatively simple and a set of bedrock terraces enables reconstruction of the palaeo-channel bed during canyon incision. Plucking is the dominant erosion mechanism in the well-jointed basalt bedrock of the Channeled Scablands<sup>6</sup>; however, forward-modelling the co-evolution of hydraulics and bed topography in a megaflood channel that is eroding by plucking is currently intractable, owing to computational expense and the fact that the physics of wall and bed erosion are not well known<sup>21</sup>. Instead, we conducted simulations for the modern Moses Coulee topography and explored how flooding evolved in concert with canyon incision by routing floods through four inferred stages of canyon formation (Fig. 1c), for which we reconstructed palaeo-topography guided by bedrock terraces (see Methods). Floods were simulated in discharge increments of  $0.1 \times 10^6 \text{ m}^3 \text{ s}^{-1}$  using a two-dimensional, depth-averaged, hydraulic model. The elevations of high-water marks identified in the field were used to constrain the discharges predicted by the brim-full model. The discharges predicted by the threshold shear stress model were determined from simulations in which bed stresses only slightly exceeded the threshold for incision via bedrock-block plucking, with lower-bound estimates based on the assumption of cohesionless blocks and upper-bound estimates that include interlocking via the inclusion of block wall friction.

<sup>1</sup>Department of Geosciences, University of Massachusetts, Amherst, Massachusetts 01003, USA. <sup>2</sup>Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA.



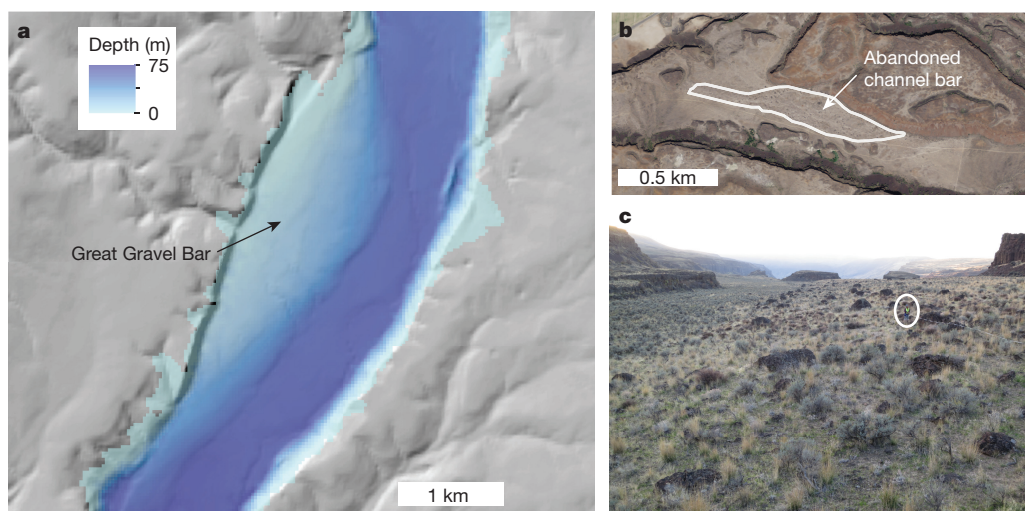
**Figure 1 | The Moses Coulee study site.** **a**, Moses Coulee in eastern Washington, USA (location indicated by the smaller boxed region in the inset). Blue areas were inundated by late-Pleistocene floods and the white line delineates the southern extent of the Cordilleran ice sheet. Flood and ice extent are interpreted from version 3.0 of the Washington Department of Natural Resources digital 1:250,000 scale state geological map. **b**, Moses Coulee with modelled bed shear stresses (see colour scale in **d**) during a flood with a discharge of  $0.6 \times 10^6 \text{ m}^3 \text{ s}^{-1}$ ; the model domain is outlined

The threshold shear stress model for bedrock canyons stems from theoretical and field evidence from gravel-bedded rivers that demonstrates that, on average, bed shear stresses can exceed the threshold for sediment transport by only 20% before the channel cross-section will erode and evolve to a form that maintains shear stresses near the threshold for sediment transport<sup>25</sup>. Hence, well-jointed rock that is subject to megaflooding may behave more like a bed of sediment with limited or no cohesion in which individual blocks can be easily entrained and transported, rather than as massive crystalline rock that erodes slowly through abrasion<sup>5,22,23</sup>. Such bedrock channels cannot withstand large discharges or shear stresses that greatly exceed the threshold for plucking. Instead, the channels will adjust their morphology through bedrock erosion so that bed stresses, on average, only slightly exceed the threshold for block erosion.

The discharges predicted by the two models are tested by assessing the temporal evolution of discharge during canyon incision, because

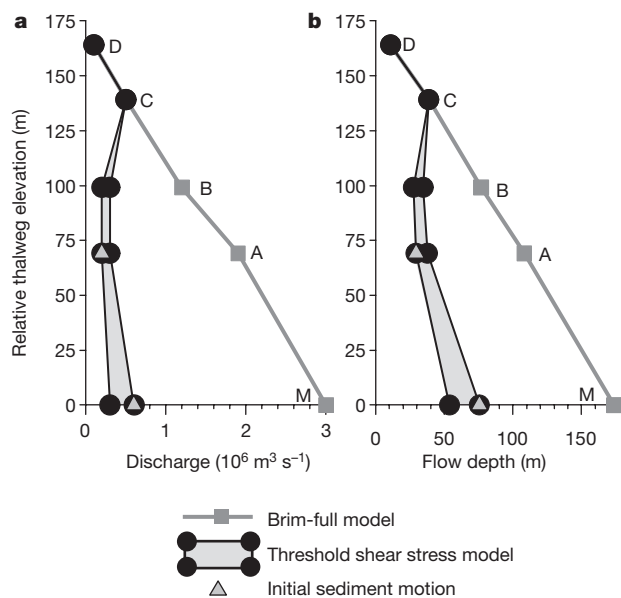
the brim-full model implies a temporal increase in discharge to maintain brim-full flow while the canyon floor erodes. However, in the Channeled Scablands, at least tens of floods are thought to have occurred<sup>24</sup>, and all but the most recent floods probably had comparable discharges, owing to a triggering mechanism that required filling an ice-dammed lake to a threshold level sufficient to float the dam and release an outburst<sup>24</sup>. A second test uses the depositional bars within Moses Coulee, because these require certain levels of inundation and sediment transport regimes to form<sup>26</sup>, by determining which model predictions are most consistent with the location and morphology of two key landforms: the Great Gravel Bar (Fig. 2a) and a boulder bar on an 'abandoned channel', a terrace 70 m above the canyon floor (Fig. 2b, c).

The discharge predicted by the brim-full model increases from  $0.1 \times 10^6 \text{ m}^3 \text{ s}^{-1}$  to a peak of  $3.0 \times 10^6 \text{ m}^3 \text{ s}^{-1}$  (Fig. 3a) during canyon formation; the peak discharge value during the final stage of canyon



**Figure 2 | Depositional landforms in Moses Coulee.** **a**, The Great Gravel Bar in upper Moses Coulee overlain by a water-depth map for a simulated flood with a discharge of  $0.6 \times 10^6 \text{ m}^3 \text{ s}^{-1}$ . **b**, Aerial photograph showing the location of the abandoned channel boulder bar (outlined in white).

Aerial imagery courtesy of the US Department of Agriculture. **c**, Field photo of the boulder deposit shown in **b**; the median grain size of clasts on the bar is 0.15 m. The photo is taken near the tip of the arrow in **b** and the view is downstream; person (circled) for scale.

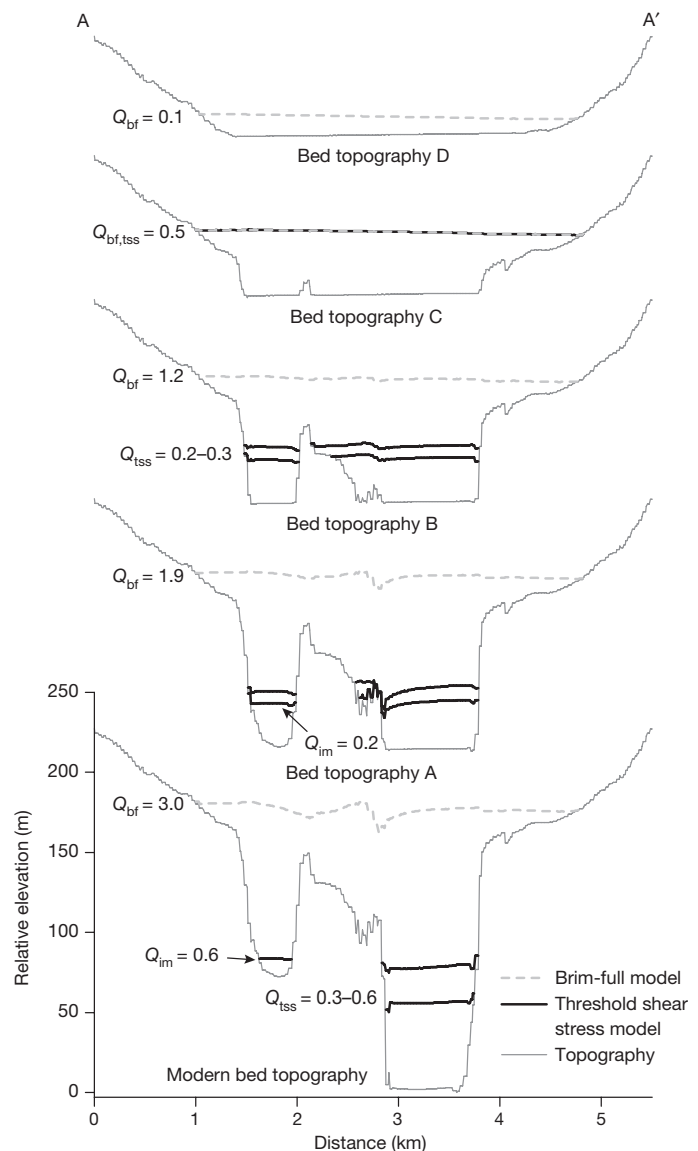


**Figure 3 | Simulated discharge and water depths.** **a, b,** Discharge (**a**) and water depths (**b**) predicted by the brim-full model, by the threshold shear stress model, and from criteria related to the initial motion of sediment. The shading shows the range of predicted values based on upper- and lower-bound parameterizations of the critical dimensionless shear stress for bedrock incision by block sliding (see Methods). The letters A–D indicate simulation results for bed elevations A–D; M denotes results for the modern topography. Relative thalweg elevations denote the elevation of the palaeo-channel floor above the modern channel floor.

formation is of similar order to previous estimates for other Channeled Scablands canyons<sup>11</sup>. By contrast, the discharges predicted by the threshold shear stress model are lower by approximately an order of magnitude, with incision from the upper-most terrace to the canyon floor achieved by discharges of at most  $0.6 \times 10^6 \text{ m}^3 \text{ s}^{-1}$  (Fig. 3a), and the predicted discharges change only slightly during canyon formation. Water depths predicted by the threshold shear stress model are 27–76 m, whereas the brim-full hypothesis predicts increasing water depths during canyon formation with a peak at 175 m (Fig. 3b). The narrow range of discharge rates predicted by the threshold shear stress model is consistent with the incision occurring during a single flood<sup>5</sup> or during multiple floods of similar magnitude<sup>9</sup>, as would be expected for floods issuing repeatedly from the same ice-dammed lake<sup>24</sup>.

The Great Gravel Bar in upper Moses Coulee is inundated at a modelled discharge of  $0.6 \times 10^6 \text{ m}^3 \text{ s}^{-1}$  (Fig. 2a), which is consistent with the upper discharge estimate predicted by the threshold shear stress model and with observations that indicate that bars tend to aggrade to near the height of the flood<sup>27</sup>. The bar at the abandoned channel site is made up of well-rounded basalt clasts with a median diameter of 0.15 m, indicating that they were transported as bedload. Bar deposition by bedload is consistent both with discharges predicted by the threshold shear stress model (because modelled bed shear stresses at the bar just exceed the threshold of motion) and with the formation regime of bars in rivers with coarse-grained bed sediment<sup>28</sup>. Shear stresses predicted by the brim-full model would have exceeded the threshold of suspension for the majority of the boulders on the bar (which have diameters of less than 0.5 m), which is inconsistent with the field evidence for bedload deposition. The morphology and sedimentology of the depositional bars are therefore more consistent with the threshold shear stress end-member than with the brim-full end-member.

Bed stresses on some parts of the knickzone in the central part of Moses Coulee are several-fold higher than the thresholds for plucking, for discharges that are consistent with the threshold shear stress model (Fig. 1b, d). If flood waters no longer reached the height of the terraces because they were abandoned by lowering of the channel bed as the



**Figure 4 | Co-evolution of topography and flood level predicted by the brim-full and threshold shear stress models.**  $Q_{bf}$ ,  $Q_{tss}$  and  $Q_{im}$  refer to discharges predicted by brim-full, threshold shear stress and sediment initial motion criteria, respectively, with units of  $10^6 \text{ m}^3 \text{ s}^{-1}$ . The location of the cross-section (A–A') is shown in Fig. 1d.

knickzone retreated upstream, as is probably the case, then the higher shear stresses on the steep channel bed<sup>29</sup> indicate that the knickzone was probably a site of transient, rapid erosion, provided that the well-jointed Columbia River basalts are indeed mechanically similar to cohesionless or weakly interlocked bedrock blocks.

The threshold shear stress model implies that canyons in the Channeled Scablands were eroded by floods with depths that were a fraction of the relief of the final canyon (Fig. 4). This physics-based finding is consistent with several recent investigations of canyon carving at other sites on Earth and Mars: for example, those where bedrock incision by plucking or toppling of jointed rock occurs at depths less than brim-full<sup>5,16–18,23</sup>, those where terrace chronology indicates multiple episodes of canyon incision<sup>9,20</sup>, or those where lakes in breached craters contain insufficient water volumes to fill downstream channels<sup>19</sup>.

Our results suggest that the morphology of canyons (for example, terraces, valley shapes and slope profiles) on Earth and Mars could reveal information about both the history and discharge of flooding that warrants further investigation. The outburst floods that carved the



Channeled Scablands were extraordinary under either end-member model, but predictions of discharges from the threshold shear stress model are five- to ten-fold smaller. On Mars, owing to the low permeability of aquifers, it has been challenging to reconcile the very large reconstructed brim-full discharges in outflow channels with a subsurface flood source<sup>30</sup>. Given the proposed similarity in incision mechanics for outflow channels on Mars and in the Channeled Scablands<sup>2,16,20,22,23</sup>, the threshold shear stress model provides a link between the physics of groundwater-sourced floods and terraces observed in orbital data<sup>20</sup>, implying longer duration, lower discharge floods, or multiple floods on early Mars<sup>30</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 22 April; accepted 31 August 2016.**

- Bretz, J. H. The Channeled Scablands of the Columbia Plateau. *J. Geol.* **31**, 617–649 (1923).
- Baker, V. R. *The Channels of Mars* (Univ. Texas Press, 1982).
- Carr, M. H. Mars: a water-rich planet? *Icarus* **68**, 187–216 (1986).
- Barber, D. et al. Forcing of the cold event of 8,200 years ago by catastrophic drainage of Laurentide lakes. *Nature* **400**, 344–348 (1999).
- Lamb, M. P. & Fongstad, M. A. Rapid formation of a modern bedrock canyon by a single flood event. *Nat. Geosci.* **3**, 477–481 (2010).
- Bretz, J. H. The Lake Missoula floods and the Channeled Scabland. *J. Geol.* **77**, 505–543 (1969).
- O'Connor, J. E., Grant, G. E. & Costa, J. E. in *Ancient Floods, Modern Hazards: Principles and Applications of Paleoflood Hydrology* (eds House, P. K. et al.) 359–385 (American Geophysical Union, 2002).
- Baker, V. et al. Ancient oceans, ice sheets and the hydrological cycle on Mars. *Nature* **352**, 589–594 (1991).
- Baynes, E. R. et al. Erosion during extreme flood events dominates Holocene canyon evolution in northeast Iceland. *Proc. Natl Acad. Sci. USA* **112**, 2355–2360 (2015).
- Carr, M. H. Formation of Martian flood features by release of water from confined aquifers. *J. Geophys. Res. Solid Earth* **84**, 2995–3007 (1979).
- Baker, V. R. *Paleohydrology and Sedimentology of Lake Missoula Flooding in Eastern Washington*. GSA Special Paper 144 (Geological Society of America, 1973).
- Denlinger, R. & O'Connell, D. Simulations of cataclysmic outburst floods from Pleistocene Glacial Lake Missoula. *Geol. Soc. Am. Bull.* **122**, 678–689 (2010).
- O'Connor, J. E. & Baker, V. R. Magnitudes and implications of peak discharges from glacial Lake Missoula. *Geol. Soc. Am. Bull.* **104**, 267–279 (1992).
- Komar, P. D. Comparisons of the hydraulics of water flows in Martian outflow channels with flows of similar scale on Earth. *Icarus* **37**, 156–181 (1979).
- Miyamoto, H. et al. Cataclysmic Scabland flooding: insights from a simple depth-averaged numerical model. *Environ. Model. Softw.* **22**, 1400–1408 (2007).
- Lamb, M. P., Mackey, B. H. & Farley, K. A. Amphitheater-headed canyons formed by megaflooding at Malad Gorge, Idaho. *Proc. Natl Acad. Sci. USA* **111**, 57–62 (2014).
- Lamb, M. P., Dietrich, W. E., Aciego, S. M., DePaolo, D. J. & Manga, M. Formation of Box Canyon, Idaho, by megaflood: implications for seepage erosion on Earth and Mars. *Science* **320**, 1067–1070 (2008).
- Anton, L., Mather, A., Stokes, M., Muñoz-Martín, A. & De Vicente, G. Exceptional river gorge formation from unexceptional floods. *Nat. Commun.* **6**, 7963 (2015).
- Coleman, N. M. Hydrographs of a Martian flood from a breached crater lake, with insights about flow calculations, channel erosion rates, and chasma growth. *J. Geophys. Res. Planets* **118**, 263–277 (2013).
- Warner, N., Gupta, S., Muller, J., Kim, J. & Lin, S. A refined chronology of catastrophic outflow events in Ares Vallis, Mars. *Earth Planet. Sci. Lett.* **288**, 58–69 (2009).
- Lamb, M. P., Finnegan, N. J., Scheingross, J. S. & Sklar, L. S. New insights into the mechanics of fluvial bedrock erosion through flume experiments and theory. *Geomorphology* **244**, 33–55 (2015).
- Lamb, M. P. & Dietrich, W. E. The persistence of waterfalls in fractured rock. *Geol. Soc. Am. Bull.* **121**, 1123–1134 (2009).
- Lapotre, M. G. A., Lamb, M. P. & Williams, R. M. E. Canyon formation constraints on the discharge of catastrophic outburst floods of Earth and Mars. *J. Geophys. Res. Planets* **121**, 1232–1263 (2016).
- Waitt, R. B. Case for periodic, colossal jökulhlaups from Pleistocene glacial Lake Missoula. *Geol. Soc. Am. Bull.* **96**, 1271–1286 (1985).
- Parker, G. Self-formed straight rivers with equilibrium banks and mobile bed. Part 2. The gravel river. *J. Fluid Mech.* **89**, 127–146 (1978).
- Wohl, E. E. Bedrock benches and boulder bars: floods in the Burdekin Gorge of Australia. *Geol. Soc. Am. Bull.* **104**, 770–778 (1992).
- Schmidt, J. C., Grams, P. E. & Leschin, M. F. in *The Controlled Flood in Grand Canyon* (eds Webb, R. H. et al.) 185–203 (American Geophysical Union, 1999).
- Church, M. Bed material transport and the morphology of alluvial river channels. *Annu. Rev. Earth Planet. Sci.* **34**, 325–354 (2006).
- Venditti, J. G. et al. Flow in bedrock canyons. *Nature* **513**, 534–537 (2014).
- Andrews-Hanna, J. C. & Phillips, R. J. Hydrological modeling of outflow channels and chaos regions on Mars. *J. Geophys. Res. Planets* **112**, E08001 (2007).

**Acknowledgements** This research was supported by a Caltech Texaco Prize Postdoctoral Fellowship, collaborative NSF (1529528, 1529110) funding to I.J.L. and M.P.L., and a NASA (NNX13AM83G) award to M.P.L. We thank E. Simon, M. Lapotre and S. Roberts for assistance and advice with the hydraulic modelling, and participants in the Caltech field methods course for field assistance.

**Author Contributions** I.J.L. and M.P.L. designed the study and wrote the manuscript. I.J.L. conducted the hydraulic simulations.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.J.L. (ilarsen@umass.edu).

**Reviewer Information** Nature thanks J. Venditti and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Hydraulic modelling.** Outburst floods through Moses Coulee were simulated using ANUGA version 2.0, a finite-volume hydrodynamic model that conserves mass and momentum by solving the two-dimensional, time-dependent, depth-averaged, shallow-water equations on a triangular mesh<sup>31,32</sup>. The size of the triangular mesh was varied spatially, with a maximum triangle area of 900 m<sup>2</sup> for the main study reach and 5,000 m<sup>2</sup> elsewhere (Extended Data Fig. 1).

Floods were simulated with five different topographic boundary conditions: the modern topography and four topographies with reconstructed bed elevations (A–D). We simulated the hydraulic conditions through inferred stages of canyon evolution that we reconstructed from terraces. The reconstructed bed elevations were based on the elevations of bedrock terraces found downstream of a prominent knickzone and maintain the same bed slope as the channel upstream from the knickzone (Fig. 1; Extended Data Fig. 2). Although we do not explicitly simulate lateral knickpoint retreat, which probably accounted for much of the erosion of Moses Coulee, the decrease in bed elevation that we impose for different simulations mimics the vertical bed lowering that would have occurred as knickpoints migrated upstream, past the reach with preserved terraces. If slopes were steeper than the modern topography during canyon incision and erosion was limited to steep reaches<sup>29</sup>, then the discharges predicted by our threshold shear stress model are upper estimates; the hydraulic simulations indicate that even the smallest floods we simulated produce shear stresses across the knickzone that exceed the threshold for block sliding (Extended Data Fig. 3).

The topographic data were from 10-m digital elevation model (DEM) data derived from US Geological Survey (USGS) topographic maps. DEM and reconstructed topographic data were used by ANUGA to generate the triangular mesh.

ANUGA implements bed friction with Manning's roughness coefficient ( $n$ ). We determine the roughness coefficient by equating a form of the Manning–Strickler relation<sup>33</sup>

$$\frac{u}{u_*} = 8.1 \left( \frac{h}{k_s} \right)^{1/6}$$

and the Manning equation

$$u = \frac{h^{2/3} S^{1/2}}{n}$$

and re-arranging for  $n$ :

$$n = \frac{1}{8.1} \frac{k_s^{1/6}}{g^{1/2}}$$

Here  $u$  is flow velocity,  $u_* = \sqrt{C_f} u$  is bed shear velocity,  $C_f$  is a friction coefficient,  $g$  is acceleration due to gravity,  $h$  is flow depth,  $S$  is the tangent of the channel-bed angle and  $k_s$  is a bed-roughness length scale. We assume that  $k_s$  follows a relation proposed for bedrock channels:

$$k_s = r_d r_{br} \sigma_{br}$$

where  $r_d$  and  $r_{br}$  are hydraulic roughness scaling parameters and  $\sigma_{br}$  is one standard deviation of the elevation of the bedrock bed<sup>34</sup>. We set  $r_d = 2$  and assume  $r_{br} = 2$ , which is within the range of reported values<sup>34</sup>. The mean of  $\sigma_{br}$  from five different reaches in our study area was about 5 m; consequently, we determined that  $n = 0.065$  and assume that this roughness coefficient is spatially uniform. The USGS 10-m DEM was used to determine roughness.

For each flood simulation, a constant discharge was input at the upstream boundary of the model domain and flow evolved within the domain before exiting at the downstream boundary<sup>32</sup>. The duration of each simulation was 10,000 s, which was sufficient for the flow to evolve to near steady-state (Extended Data Fig. 4). Model outputs for flood elevation (stage) and the  $x$  and  $y$  components of momentum and velocity were saved every 100 s and gridded to a pixel resolution of 30 m  $\times$  30 m, a value selected because it is comparable to the maximum triangle area in the study reach. The 10 grids from the final 1,000 s of each simulation were averaged. From the averaged grids, flow depth ( $h$ ) was calculated as the difference in mean flood stage and bed topography, streamwise velocity ( $u$ ) was determined by the vector sum of the  $x$  and  $y$  velocity components, and bed shear stress ( $\tau_b$ ) was calculated as  $\tau_b = \rho C_f u^2$ , where  $\rho$  is the density of water. The friction coefficient is related to Manning's  $n$  by

$$C_f = \frac{g n^2}{h^{1/3}}$$

**Discharge predicted by the brim-full model.** The discharge predicted by the brim-full model was determined by iteratively simulating floods in discharge

increments of  $0.1 \times 10^6 \text{ m}^3 \text{ s}^{-1}$  through the modern topography and identifying the discharge scenario in which the simulated flood stage reached the brim-full level determined from field evidence. Field evidence for the elevation of the brim-full level was a scarp cut into loess with rounded basalt clasts deposited at its base.

**Discharge predicted by the threshold shear stress model.** Theoretical and field evidence from gravel-bedded rivers demonstrates that bed shear stresses can exceed the threshold for sediment transport by a factor of only about 1.2 on average before the channel cross-section erodes and evolves to a form that maintains shear stresses near the threshold for sediment transport<sup>25</sup>. Ultimately, to form a stable alluvial channel, the channel banks must be below the threshold for erosion, whereas the bed must be above the threshold for transport. If the bed and banks are composed of the same material, which is the case both for gravel-bed rivers and the Channeled Scablands, then the channel-forming shear stress for self-formed channels cannot greatly exceed the conditions required for sediment entrainment. The similarity in low excess shear stress (factor of 1.1–1.6) required to erode gravel-bedded channels and the channel at Canyon Lake Gorge, Texas<sup>5</sup>, suggests that the mechanics of erosion and sediment transport in gravel-bedded rivers are similar to those in bedrock channels formed in well-jointed bedrock, where plucking of bedrock blocks is the dominant erosion mechanism<sup>6,11,21</sup>, leading to the threshold shear stress end-member model for canyon formation. A similar threshold stress mechanism has also been proposed at waterfalls dominated by toppling, which are common to Scabland terrain<sup>23</sup>.

The threshold shear stress model predicts that channel bed stresses should be at or near the threshold for plucking, but local deviations are observed (Fig. 1) and expected for a variety of reasons, consistent with observations in threshold alluvial rivers. Anomalies are likely to occur where areas of the bed are eroding faster, such as in knickzones, in areas where heterogeneity in rock strength both within and among basalt flows leads to differing fracture spacing and block size (leading to, for example, differences in canyon width), and in areas where the formation of backwaters locally lowers bed stresses. For example, our simulations assume a constant block size of 0.5 m; accounting for a wider range of block sizes would produce an equally wider range of expected threshold stresses for plucking (equation (1)). The threshold shear stress model predicts that, over time, the topography is likely to evolve such that bed protrusions are eroded and knickpoints are transient and retreat, such that shear stresses, on average, tend towards the threshold values for erosion. For Moses Coulee, 67% of the terrain has simulated bed stresses within the bounds for plucking blocks of the size found in depositional bars in the Channeled Scablands (0.13–0.83 m), which is consistent with the threshold shear stress model; for brim-full flow, only 33% of the terrain has simulated bed stresses within the plucking bounds for the same range of block sizes (Extended Data Fig. 5).

Plucking—the removal of bedrock blocks—is the dominant erosion mechanism in the well-jointed basalt bedrock of the Channeled Scablands<sup>6,22</sup>. Experimental studies have shown that downstream sliding of blocks can be the dominant form of plucking<sup>35</sup>. Therefore, to evaluate the discharges predicted by the threshold shear stress model, we first determined the dimensionless critical shear stress for block sliding:

$$\tau_{pc}^* = \frac{\cos(\theta)[\tan(\phi) - \tan(\theta)] + 2\tau_w^*}{\left[1 + \frac{1}{2}C_D \left(\frac{u}{u_*}\right)^2 \frac{P}{L}\right] [1 + F_L^* \tan(\phi)]}$$

where  $\theta$  is the bed angle,  $\phi$  is the bed friction angle,  $\tau_w^*$  is the dimensionless block sidewall stress,  $C_D$  is the local drag coefficient,  $P$  is the block protrusion height (or equivalently the roughness on the top of the block),  $L$  is the block length and  $F_L^*$  is the dimensionless hydraulic lift force<sup>21</sup>. We set  $\theta = 1^\circ$  on the basis of measurements of the topographic slope outside of the channel upstream from the terraces in Moses Coulee. The friction angle was assumed to be  $34^\circ$ , consistent with the range of values ( $31^\circ$ – $36^\circ$ ) for wet basalt<sup>36</sup>. The ratio of block protrusion height to block length ( $P/L$ ) was set to 0.2 on the basis of field observations. The quantities  $C_D$ ,  $u/u_*$  and  $F_L^*$  were taken to be 1, 8.3 and 0.85, respectively<sup>21</sup>. Dimensionless block sidewall stress is poorly constrained, so we varied the values between 0 and 0.2 to predict a range of potential values for  $\tau_{pc}^*$ . The Columbia River basalts are jointed and fractured, commonly exhibiting colonnade and entablature structures<sup>37</sup>. The fractures reduce the cohesive strength of the Columbia River basalts by one or two orders of magnitude relative to intact basalt<sup>38</sup>. A dimensionless block sidewall stress of 0 is appropriate for the wide, open joints we observed between basalt columns while in the field, and 0.2 corresponds to interlocking or cohesion between blocks with a wall stress that is 20% of the block weight per unit area. Greater assumed cohesive strength or interlocking would lead to higher discharge estimates for the threshold shear stress model. The range of  $\tau_{pc}^*$  was then used to calculate upper and lower bounds on the critical shear stress for block sliding:

$$\tau_{pc} = \tau_{pc}^* (\rho_s - \rho) g D \quad (1)$$

where  $\rho_s$  is the density of the basalt ( $2,800 \text{ kg m}^{-3}$ ) and  $D$  is the block height<sup>21</sup>. We set  $D = 0.5 \text{ m}$ , a value consistent with the dimensions of basalt columns<sup>39</sup> in the Channeled Scablands and that falls within the range of median intermediate-axis diameters ( $D_{50}$ ) of boulders we measured on the Ephrata Fan ( $D_{50} = 0.30 \text{ m}$ , sample size  $N = 138$ ) and in Drumheller Channels ( $D_{50} = 0.59 \text{ m}$ ,  $N = 301$ ) (Extended Data Fig. 6). The Ephrata Fan and Drumheller Channels sites are located in a different floodway approximately 30 km and 70 km to the southeast of Moses Coulee, respectively<sup>6</sup>. The lower and upper shear stress bounds for erosion by sliding calculated by this method (based on varying the end-member block sidewall stress values from 0 to 0.2) are 467 Pa and 751 Pa, respectively.

Unlike the brim-full discharge model, which predicts discharge for the length of an entire channel reach, the threshold shear stress model makes predictions for local bed shear stress within the channel. We therefore determined the mean and standard deviation of shear stress at 12 locations placed approximately along a cross-section within the study reach, which span an elevation range from the channel floor to the highest terrace (Extended Data Fig. 7). To determine the discharges predicted by the threshold shear stress model, we used the upper and lower shear stress bounds for plucking by sliding shown in Extended Data Fig. 8 to define the widest possible bounds on flood discharge that satisfy the model conditions for all eroding sites.

For a given topographic reconstruction, corresponding to the modern topography or to one of our four inferred stages of canyon incision, we first used the upper and lower thresholds for plucking to define the range of possible discharges at the lowest elevation site that are consistent with the threshold shear stress model. For the modern topography, for example, the lowest site corresponds to the modern canyon floor (Extended Data Fig. 7). We then found the discharge range, following the same procedure, at sites with progressively higher elevations that produced bed stresses that were within the bounds for plucking at the site of interest and also for all lower elevation sites (Extended Data Fig. 9). These two criteria are important because, for a flood discharge to be consistent with the threshold stress model, all inundated and eroding sites from the canyon floor upward must have bed stresses that are within the bounds for plucking. Most brim-full discharge scenarios violate these constraints because they produce bed stresses on the canyon floor that greatly exceed the bounds on plucking (Extended Data Fig. 8). For the threshold shear stress model, if the discharge required to pluck blocks at a higher elevation site exceeds that of a lower elevation site, then the higher sites were considered to be dry and abandoned and therefore were not used to define the range of discharges for that stage of canyon incision (Extended Data Fig. 9). This analysis therefore allows for identification, for a given topographic reconstruction, of the maximum range in flood discharges that is consistent with the threshold shear stress model, for both the canyon floor and all higher elevation surfaces that could have been inundated. The procedure was repeated for the modern topography and the topographies for each of the four stages of canyon incision to define discharges, flow depths, cross-sectional areas and channel widths (Fig. 3, Extended Data Fig. 10).

**Thresholds for sediment transport and suspension.** The abandoned channel on the south side of the study reach contains a gravel-boulder bar with well-rounded

clasts (Fig. 2b, c). We measured the intermediate diameter of 220 individual clasts measured at 1-m spacing along a grid (excluding loess deposited after flooding) and calculated  $D_{50} = 0.15 \text{ m}$  and  $D_{84} = 0.41 \text{ m}$  (84th percentile of clast diameters are finer than this size) (Extended Data Fig. 6). We calculated the critical shear stress for initial sediment motion ( $\tau_b$ ) by substituting  $\tau_c^* = 0.045$  for  $\tau_{pc}^*$  and  $D_{50}$  for  $D$  in equation (1). The resulting critical shear stress for initial motion is 119 Pa. The threshold for sediment suspension was calculated as<sup>40</sup>  $\tau_b = \rho(0.4w_s)^2$ , where  $w_s$ , the terminal settling velocity, is defined as

$$w_s = \frac{RgD^2}{C_1\nu + [0.75C_2RgD^3]^{0.5}}$$

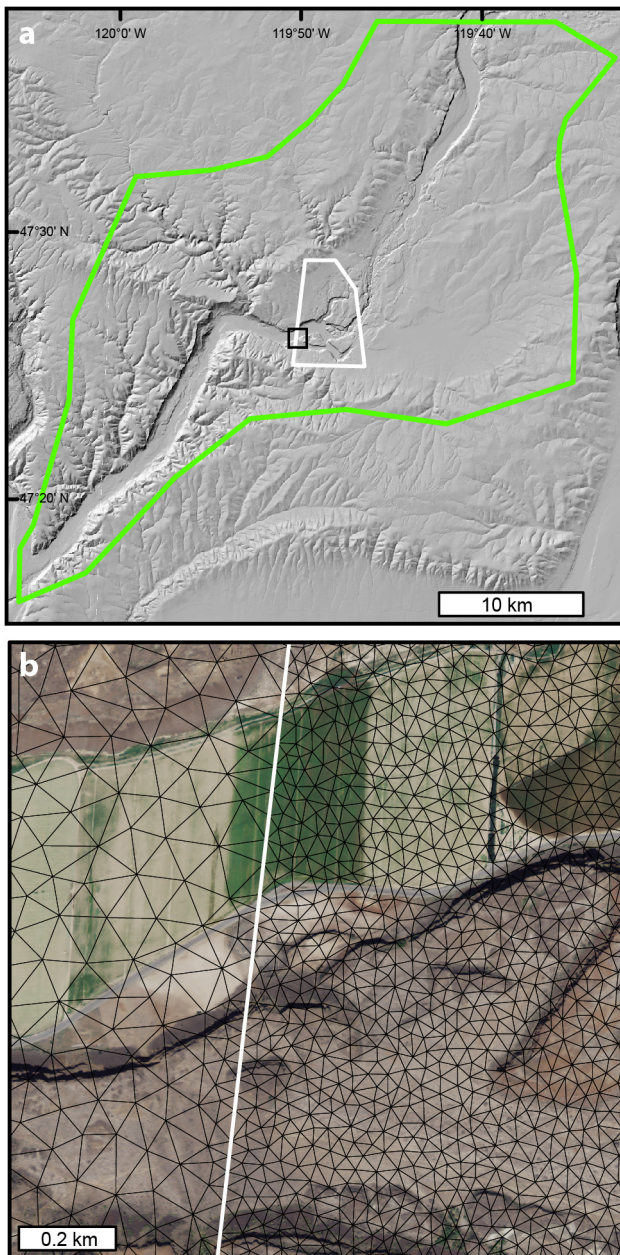
in which  $R$  is submerged specific gravity (1.8 for basalt in water),  $C_1 = 20$  and  $C_2 = 1.1$  are constants representing natural sediment, and  $\nu$  is the kinematic viscosity of water<sup>41</sup>. The shear stress threshold for suspension of  $D_{50}$  is 513 Pa and for  $D = 0.5 \text{ m}$  is 1,712 Pa, which is similar to the modelled brim-full shear stress at the bar for the modern topography (1,706 Pa), but inconsistent with bar deposition by bedload transport.

**Code availability.** The hydrodynamic code ANUGA is open-source and available for download at <https://anuga.anu.edu.au/>. The PYTHON scripts used to implement ANUGA are available from the authors by request.

**Data availability.** Digital elevation data are available from the University of Washington Geomorphological Research Group website (<http://gis.ess.washington.edu/data/>). All simulation results and data are available from the authors by request.

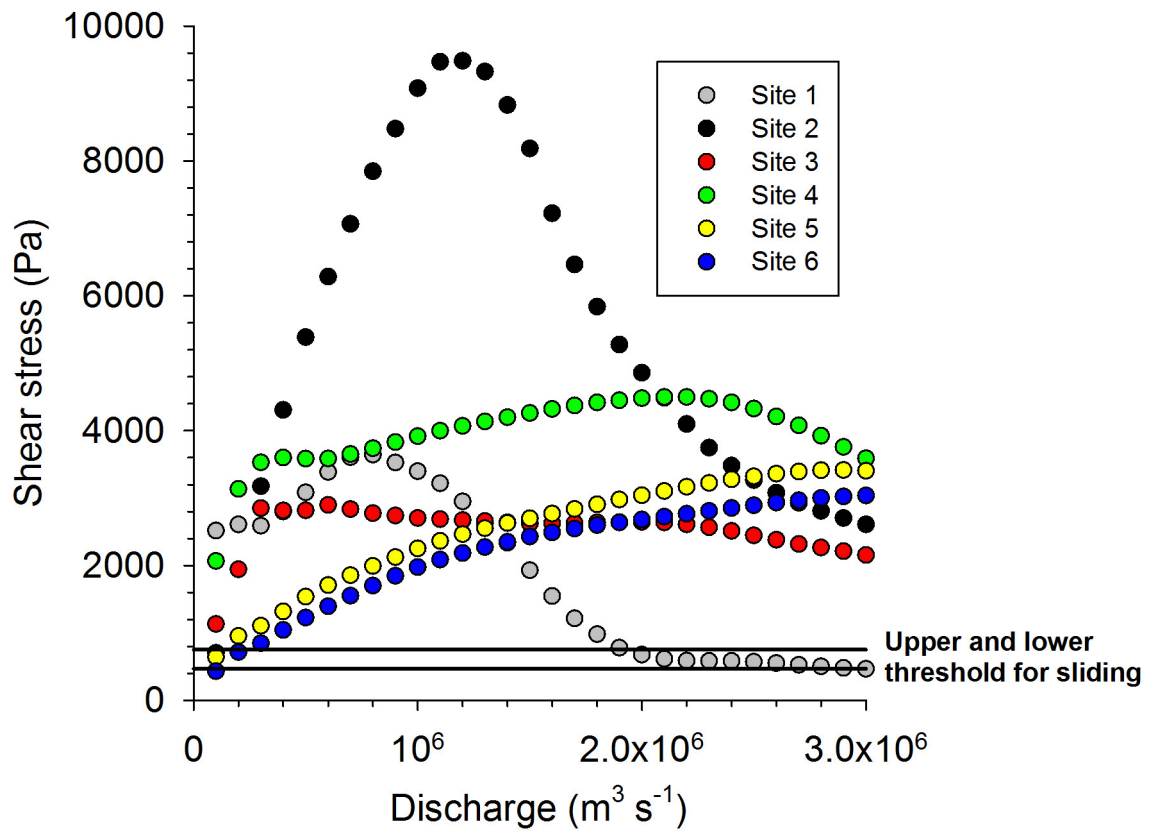
31. Roberts, S. G., Nielsen, O. M. & Jakeman, J. in *Modeling, Simulation and Optimization of Complex Processes* (eds Bock, H.G. et al.) 489–498 (Springer, 2008).
32. Roberts, S., Nielsen, O., Gray, D., Sexton, J. & Davies, G. *ANUGA User Manual. Release 2.0*. [https://github.com/GeoscienceAustralia/anuga\\_core/raw/master/doc/anuga\\_user\\_manual.pdf](https://github.com/GeoscienceAustralia/anuga_core/raw/master/doc/anuga_user_manual.pdf) (2015).
33. Parker, G. Selective sorting and abrasion of river gravel. II: applications. *J. Hydraul. Eng.* **117**, 150–171 (1991).
34. Johnson, J. P. A surface roughness model for predicting alluvial cover and bed load transport rate in bedrock channels. *J. Geophys. Res. Earth Surface* **119**, 2147–2173 (2014).
35. Dubinski, I. M. & Wohl, E. Relationships between block quarrying, bed shear stress, and stream power: a physical model of block quarrying of a jointed bedrock channel. *Geomorphology* **180–181**, 66–81 (2013).
36. Selby, M. J. *Hillslope Materials and Processes* 93 (Oxford Univ. Press, 1993).
37. Long, P. E. & Wood, B. J. Structures, textures, and cooling histories of Columbia River basalt flows. *Geol. Soc. Am. Bull.* **97**, 1144–1155 (1986).
38. Schultz, R. Limits on strength and deformation properties of jointed basaltic rock masses. *Rock Mech. Rock Eng.* **28**, 1–15 (1995).
39. Ehlmann, B. L., Viles, H. A. & Bourke, M. C. Quantitative morphologic analysis of boulder shape and surface texture to infer environmental history: a case study of rock breakdown at the Ephrata Fan, Channeled Scabland, Washington. *J. Geophys. Res. Earth Surface* **113**, F02012 (2008).
40. Niño, Y., Lopez, F. & Garcia, M. Threshold for particle entrainment into suspension. *Sedimentology* **50**, 247–263 (2003).
41. Ferguson, R. & Church, M. A simple universal equation for grain settling velocity. *J. Sediment. Res.* **74**, 933–937 (2004).





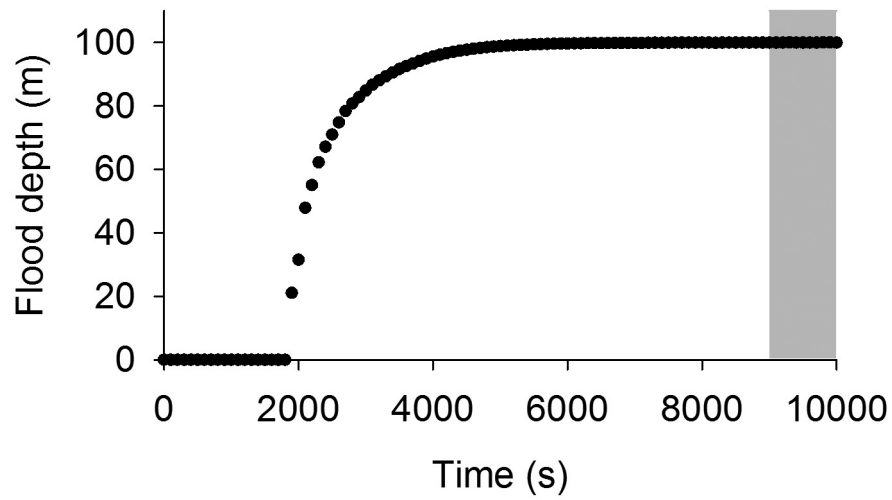
**Extended Data Figure 1 | Computational mesh resolution.** **a**, Map of the computational domain (green outline). Black square shows the location of **b**. **b**, Map showing the different triangular mesh resolutions within the computational domain, with a maximum triangle area of  $900 \text{ m}^2$  within the white polygon in **a** and  $5,000 \text{ m}^2$  throughout the rest of the domain. We used a smaller triangle area in the study reach (within the white polygon) to better resolve spatial variability in shear stresses and a slightly larger triangle area elsewhere for computational efficiency.



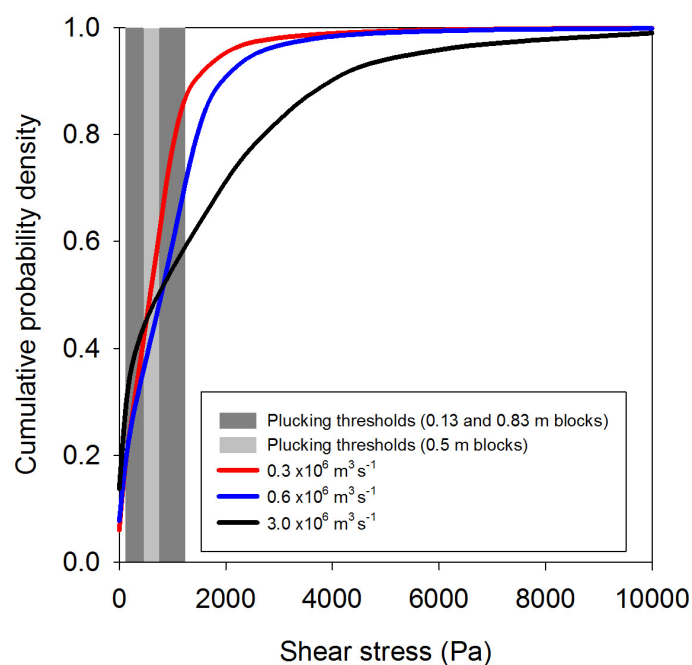


**Extended Data Figure 3 | Bed shear stress as a function of discharge for sites on the Moses Coulee knickzone.** Data correspond to locations shown in Extended Data Fig. 7. For nearly all simulated discharges, the modelled shear stresses on the knickzone exceed the thresholds for plucking via block sliding, indicating that the knickzone would probably have been rapidly eroding during floods or that it is made up of stronger rock.

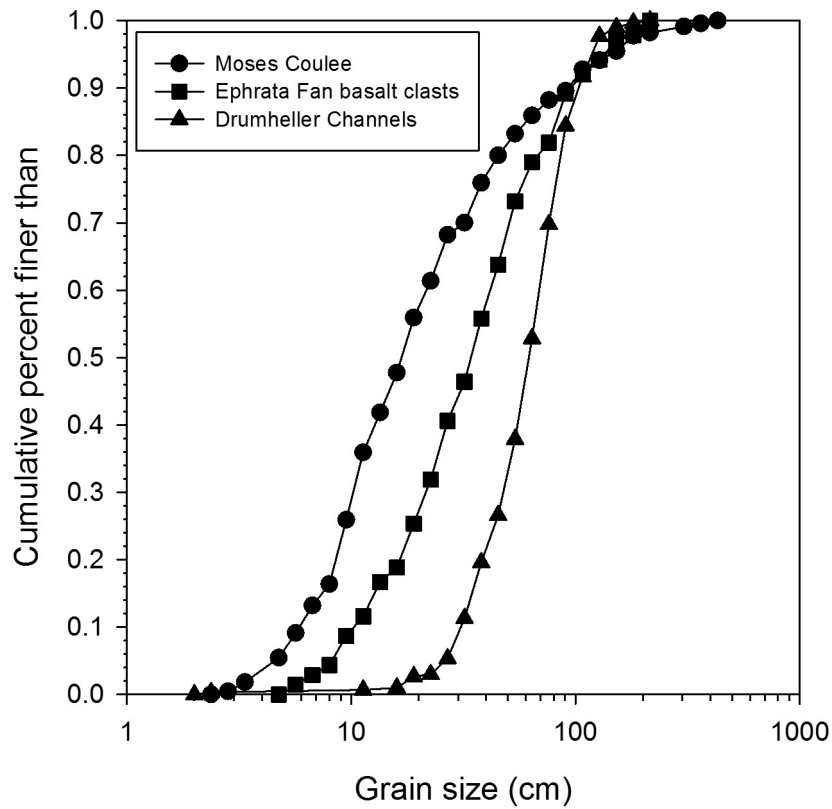




**Extended Data Figure 4 | Flood depth versus time for a typical simulation.** The example shown is a flood with a discharge of  $10^6 \text{ m}^3 \text{ s}^{-1}$  routed through the modern topography. The grey box shows the final ten time-steps, from which the model results (stage, velocity, shear stress and so on) were extracted and used to produce time-averaged grids to reduce the influence of transient waves relative to data from a single time-step.



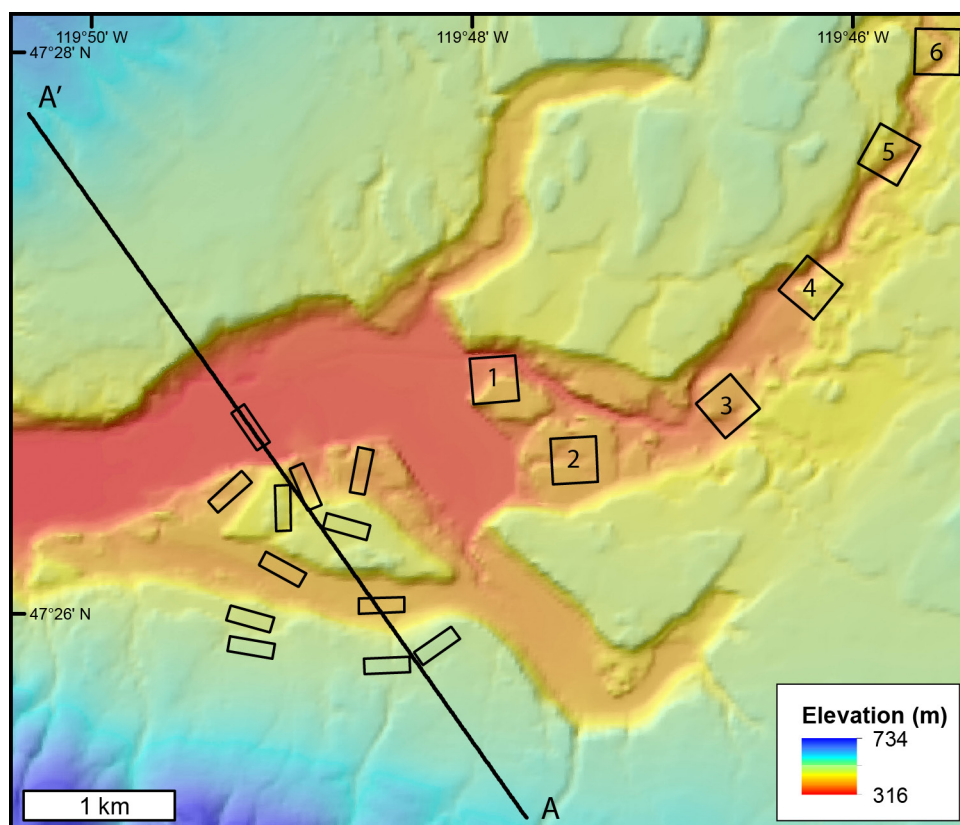
**Extended Data Figure 5 | Cumulative bed stress probability density functions.** The light grey bar shows the lower (467 Pa) and upper (751 Pa) threshold shear stress bounds for plucking of 0.5-m blocks. The dark grey bar shows the threshold stress bounds (117 Pa and 1,242 Pa) assuming a wider distribution of block sizes of 0.13–0.83 m, based on the  $D_{16}$  and  $D_{84}$  (16th and 84th percentiles) of basalt clasts at the Ephrata Fan and Drumheller Channels sites, respectively (Extended Data Fig. 6). Cumulative probability density functions are shown for the brim-full flood (discharge of  $3.0 \times 10^6 \text{ m}^3 \text{ s}^{-1}$ ; black line) and the lower ( $0.3 \times 10^6 \text{ m}^3 \text{ s}^{-1}$ ; red line) and upper ( $0.6 \times 10^6 \text{ m}^3 \text{ s}^{-1}$ ; blue line) flood discharge bounds predicted by the threshold shear stress model for the modern Moses Coulee topography. Data are extracted from each grid cell within the entire length of Moses Coulee (Fig. 1b). For the brim-full flood, only 8% of the terrain of Moses Coulee has modelled bed shear stresses within the bounds for plucking 0.5-m blocks, and this increases to 33% when considering the wider range of block sizes; some areas have extremely high bed stresses of more than 10,000 Pa. In contrast, for discharges predicted by the threshold shear stress model, 25% of the terrain has shear stresses within the bounds for plucking 0.5-m blocks, and this increases to 67% when considering a wider distribution of block sizes. This high proportion of bed stresses within the plucking threshold range, relative to the brim-full flood, is consistent with the hypothesis that the channel adjusts so that a large portion of the terrain is near the threshold state during megaflood incision. Note that the study locations in Extended Data Fig. 7 were used to define the discharge bounds for the threshold shear stress model, so all of those locations have bed stresses within the plucking threshold range by definition (see Extended Data Fig. 8).



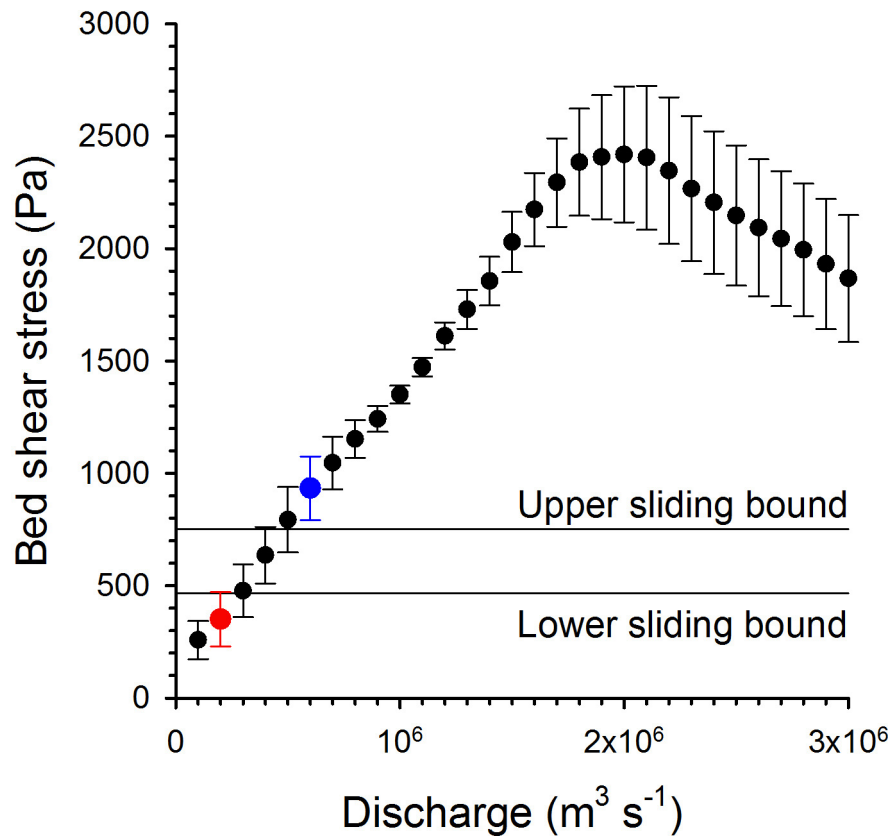
**Extended Data Figure 6 | Boulder size data.** Cumulative grain-size distribution for the Moses Coulee abandoned channel boulder bar (circles), Ephrata Fan basalt clasts (squares) and the Drumheller Channels boulder bar (triangles). Measurements were made of the intermediate grain diameter using a Wolman-style pebble count. The Ephrata Fan and

Drumheller Channels sites are to the southwest of Moses Coulee and in a different flood pathway. The larger boulders at the Ephrata Fan and Drumheller Channels are assumed to originate as basalt columns that have been rounded by fluvial transport.



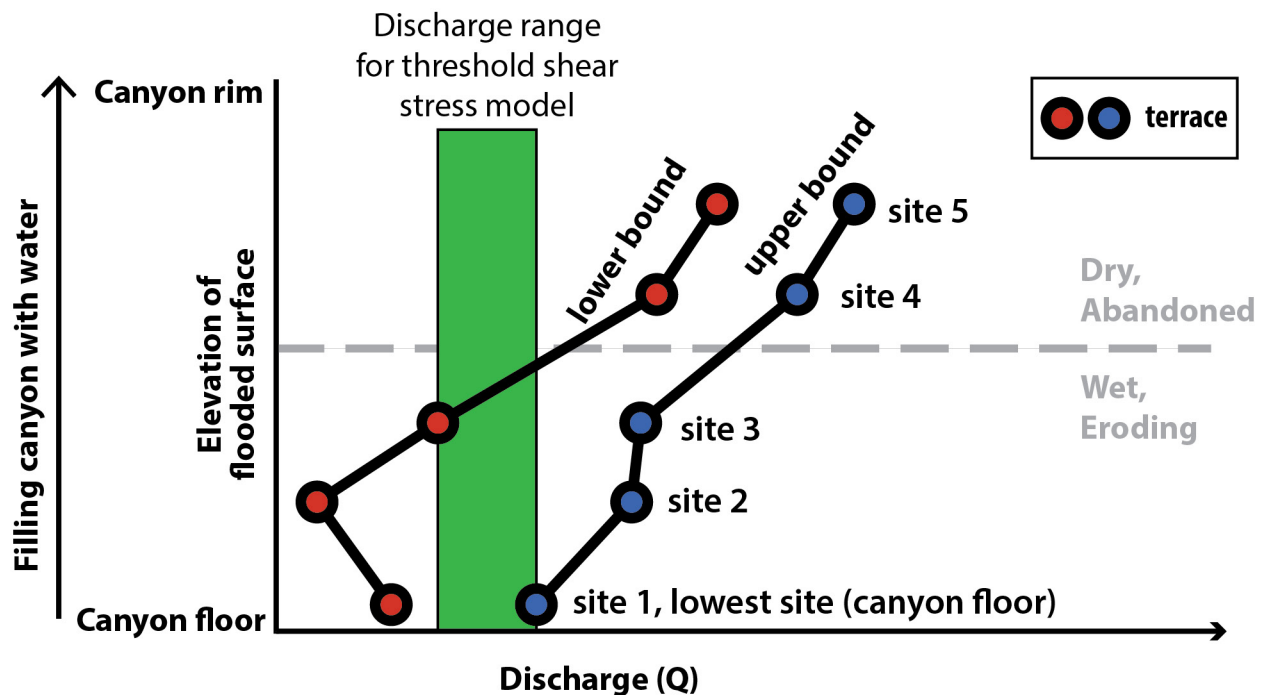


**Extended Data Figure 7 | Locations of the areas used to calculate mean bed shear stress.** The 30,000-m<sup>2</sup> rectangles near the A–A' cross-section were used to constrain discharges predicted by the threshold shear stress model. The larger 90,000-m<sup>2</sup> polygons (labelled 1–6) were used to determine the shear stresses on the knickzone; the numbers correspond to the data in Extended Data Fig. 3.



**Extended Data Figure 8 | Example model output showing the upper and lower shear stress thresholds for block sliding.** Round symbols depict the mean and bars depict one standard deviation ( $n = 33$ ) of bed shear stress values extracted from one of the twelve 30,000- $\text{m}^2$  polygons shown in Extended Data Fig. 7. In this example, the lower discharge bound

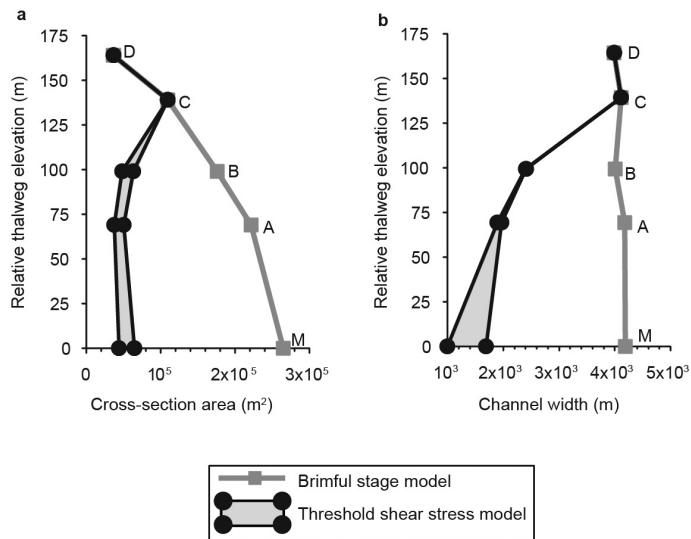
(red symbol) for the threshold shear stress model is defined as the lowest discharge for which modelled shear stresses exceed the lower shear stress threshold for sliding. The high discharge bound (blue symbol) is defined as the lowest discharge with modelled shear stresses that exceed the upper shear stress threshold for sliding.



**Extended Data Figure 9 | Schematic illustrating the method for defining the discharge range predicted by the threshold shear stress model.** For a given topographic reconstruction we first used the upper and lower bed stress thresholds for block plucking via sliding to define the range of possible discharges that are consistent with the threshold stress model at the lowest elevation site (Extended Data Fig. 7). The lowest elevation was always the canyon floor. We then followed the same procedure to determine the discharge range at sites with progressively higher elevations (for example, sites 2–5) that produced bed stress that were within the bounds for plucking at the site of interest and for all lower

elevation sites. The range of discharges determined by these two criteria is shown by the green box. These two criteria are important because, for a flood discharge to be consistent with the threshold shear stress model, all sites that are inundated and eroding, from the canyon floor upward, must have bed stresses that are within the bounds for plucking. If the discharge required to pluck blocks at a higher elevation site exceed that of a lower elevation site, then the higher sites were considered to be dry and abandoned and so were not used to define the range of discharges for that stage of canyon incision.





**Extended Data Figure 10 | Predicted channel cross-sectional areas and widths. a, b,** Cross-sectional area (a) and channel width (b) predicted by the brim-full model and threshold shear stress model. The shading shows the range of predicted values based on upper- and lower-bound parameterizations of the critical dimensionless shear stress for bedrock incision by block sliding (see Methods). For the initial topography (bed elevation D), where basalt is primarily overlain by loess, modelled shear stresses are lower than the threshold for plucking. At bed elevation C, modelled shear stresses at brim-full discharge are within 5% of the lower threshold for plucking; hence, bed stresses are assumed to be sufficient for plucking when flow was brim-full. For all other bed topographies, brim-full discharge greatly exceeds predicted values for plucking and sediment initial motion. The letters A–D denote simulation results for bed elevations A–D; M denotes results for the modern topography.

# The phylogenetic roots of human lethal violence

José María Gómez<sup>1,2</sup>, Miguel Verdú<sup>3</sup>, Adela González-Megías<sup>4</sup> & Marcos Méndez<sup>5</sup>

**The psychological, sociological and evolutionary roots of conspecific violence in humans are still debated, despite attracting the attention of intellectuals for over two millennia<sup>1–11</sup>. Here we propose a conceptual approach towards understanding these roots based on the assumption that aggression in mammals, including humans, has a significant phylogenetic component. By compiling sources of mortality from a comprehensive sample of mammals, we assessed the percentage of deaths due to conspecifics and, using phylogenetic comparative tools, predicted this value for humans. The proportion of human deaths phylogenetically predicted to be caused by interpersonal violence stood at 2%. This value was similar to the one phylogenetically inferred for the evolutionary ancestor of primates and apes, indicating that a certain level of lethal violence arises owing to our position within the phylogeny of mammals. It was also similar to the percentage seen in prehistoric bands and tribes, indicating that we were as lethally violent then as common mammalian evolutionary history would predict. However, the level of lethal violence has changed through human history and can be associated with changes in the socio-political organization of human populations. Our study provides a detailed phylogenetic and historical context against which to compare levels of lethal violence observed throughout our history.**

Debate on the nature of human violence has been ongoing since before the publication of *Leviathan* by Thomas Hobbes in 1651. Lethal violence is considered by some to be mostly a cultural trait<sup>5,6,12</sup>; however, aggression in mammals, including humans<sup>13,14</sup>, also has a genetic component with high heritability. Consequently, it is widely acknowledged that evolution has also shaped human violence<sup>2–4</sup>. From this perspective, violence can be seen as an adaptive strategy, favouring the perpetrator's reproductive success in terms of mates, status or resources<sup>15,16</sup>. Yet this does not mean that violence is invariant or even adaptive in all situations<sup>15</sup>. In fact, given that the conditions under which violence benefits evolutionary fitness depend on the ecological and cultural context, levels of violence tend to vary among human populations<sup>12,13,15,16</sup>. Disentangling the relative importance of cultural and non-cultural components of human violence is challenging<sup>3,5</sup> owing to the complex interactions between ecological, social, behavioural and genetic factors.

Conspecific violence is not exclusive to humans. Many primates exhibit high levels of intergroup aggression and infanticide<sup>4,10</sup>. Social carnivores sometimes kill members of other groups and commit infanticide when supplanting older members of the same group<sup>17,18</sup>. Even seemingly peaceful mammals such as hamsters and horses sometimes kill individuals of their own species<sup>19,20</sup>. The prevalence of aggression throughout Mammalia raises the question of the extent to which levels of lethal violence observed in humans are as expected, given our position in the phylogenetic tree of mammals. In this study, we quantified the level of lethal violence in 1,024 mammalian species from 137 families (Supplementary Information section 9a) and in over

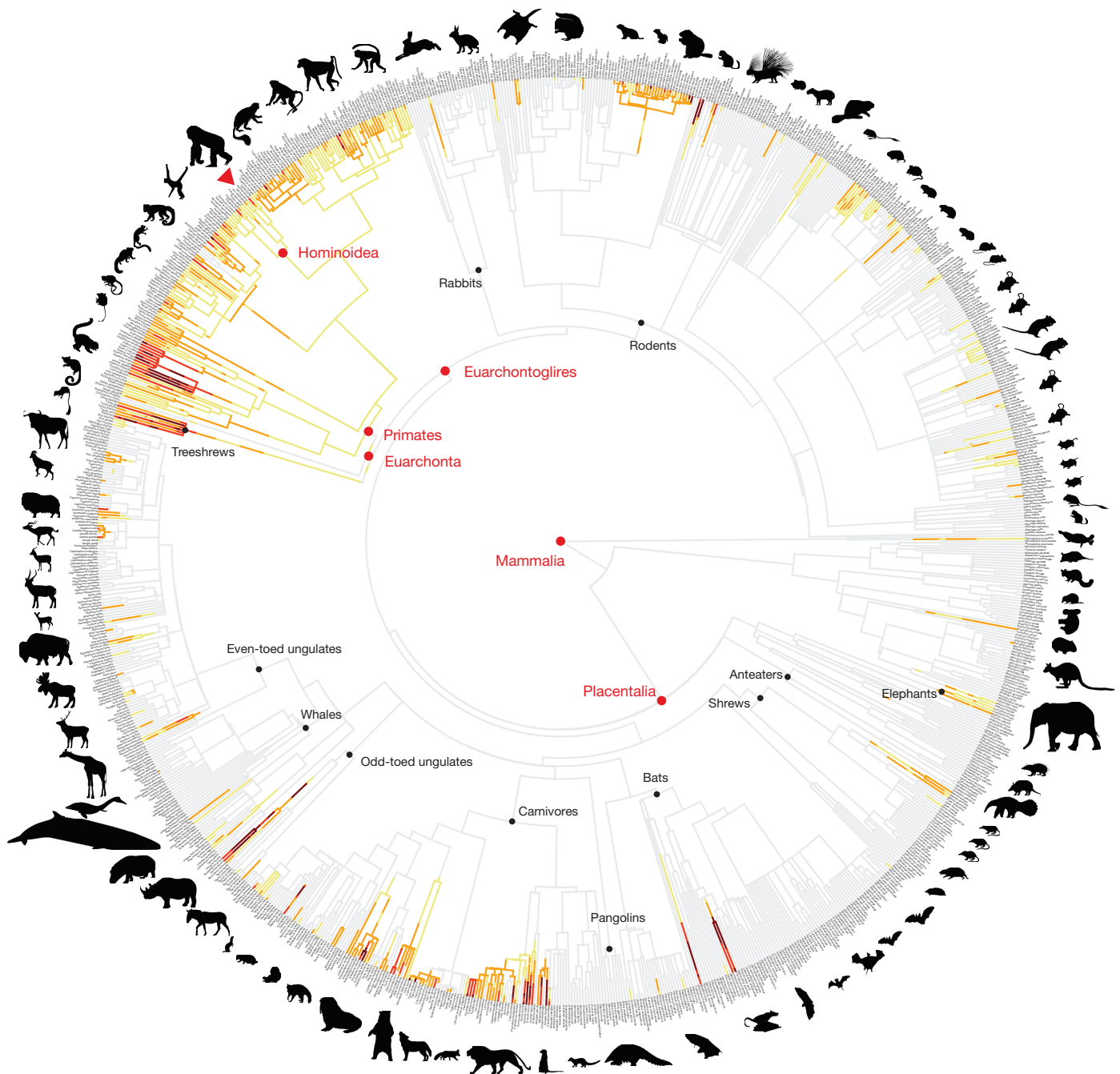
600 human populations, ranging from the Palaeolithic era to the present (Supplementary Information section 9c). The level of lethal violence was defined as the probability of dying from intraspecific violence compared to all other causes. More specifically, we calculated the level of lethal violence as the percentage, with respect to all documented sources of mortality, of total deaths due to conspecifics (these were infanticide, cannibalism, inter-group aggression and any other type of intraspecific killings in non-human mammals; war, homicide, infanticide, execution, and any other kind of intentional conspecific killing in humans).

Lethal violence is reported for almost 40% of the studied mammal species (Supplementary Information section 9a). This is probably an underestimation, because information is not available for many species. Overall, including species with and without lethal violence, we found that the percentage of deaths due to conspecifics was  $0.30 \pm 0.19\%$  of all deaths (phylogenetically corrected mean  $\pm$  s.e.m.). This level was not affected by the number of individuals sampled per species (Supplementary Information section 1). These findings suggest that lethal violence, although infrequent, is widespread among mammals<sup>19–21</sup>.

We determined whether related species tended to have similar levels of lethal violence by calculating the phylogenetic signal. We used the most recently updated mammalian phylogenies, including 5,020 extant mammals<sup>22</sup> and 5,747 extant and recently extinct mammals<sup>23</sup>. We found a significant phylogenetic signal for lethal violence, even after combining disparate causes of intraspecific killings ( $\lambda > 0.60$ ,  $P < 0.0001$ ; Supplementary Information section 2). While lethal violence was uncommon in certain clades such as bats, whales and lagomorphs, it was frequent in others, such as primates (Fig. 1). The phylogenetic signal was also significantly lower than one ( $P < 0.0001$ ), indicating that lethal violence exhibits certain evolutionary flexibility (Fig. 1). For example, the level of lethal violence strongly differs between chimpanzees (*Pan troglodytes*) and bonobos (*Pan paniscus*)<sup>10,17,20</sup>. This outcome suggests that additional factors may subsequently modify the level of lethal violence in related species. Territoriality and social behaviour mediate conspecific aggression in mammals<sup>20,24</sup>. We scored these two traits for every mammal in our study and statistically related them to the level of lethal violence using phylogenetic generalized linear models. Using this method, we found that the level of lethal violence was higher in social and territorial species than in solitary and non-territorial species (Fig. 2; Extended Data Table 1).

The occurrence of a phylogenetic signal for lethal violence in mammals enables the phylogenetic inference of lethal violence in humans. We used ancestral-state estimation methods that infer the value of a trait in any extant species according to its position in the phylogenetic tree<sup>25</sup>. The level of human lethal violence was estimated both with and without considering the territoriality and sociability of mammals. Because phylogenetic inferences are much more accurate and reliable when including information from close relatives<sup>26</sup>

<sup>1</sup>Estación Experimental de Zonas Áridas (EEZA-CSIC), E-04120 Almería, Spain. <sup>2</sup>Dpto de Ecología, Universidad de Granada, E-18071 Granada, Spain. <sup>3</sup>Centro de Investigaciones sobre Desertificación (CSIC-UV-GV), E-46113 Valencia, Spain. <sup>4</sup>Dpto de Zoología, Universidad de Granada, E-18071 Granada, Spain. <sup>5</sup>Área de Biodiversidad y Conservación, Universidad Rey Juan Carlos, E-28933 Madrid, Spain.



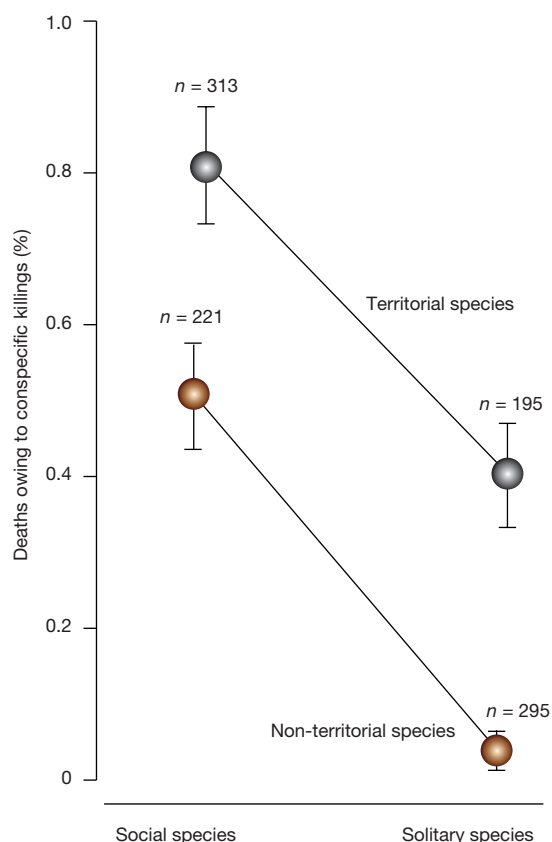
**Figure 1 | Evolution of lethal aggression in non-human mammals.** Tree showing the phylogenetic estimation of the level of lethal aggression in mammals ( $n = 1,024$  species) using stochastic mapping. Lethal aggression increases with the intensity of the colour, from yellow to dark red. Light grey indicates the absence of lethal aggression. Mammalian ancestral nodes compared with human lethal violence are shown in red, whereas main placental lineages are marked with black nodes. The red triangle indicates the phylogenetic position of humans. The silhouettes of representative mammals (downloaded from

<http://www.phylopic.org>) illustrate the main mammalian clades. They are licenced for use in the Public Domain without copyright, except for the silhouettes of Murinae (D. Liao), *Jaculus* (M. Karaka), *Philander* (S. Werning), *Rattus* (R. Groom), *Molossus* (Zimices), *Balaenoptera* (C. Hoh), *Rousettus* (O. Peles), *Connochaetes*, *Redunca*, and *Kobus* (J. A. Venter, H. H. T. Prins, D. A. Balfour and R. Slotow), that are licenced under a Creative Commons 3.0 license (<http://creativecommons.org/licenses/by/3.0>).

and fossils<sup>23</sup>, information on *Homo neanderthalensis* was included when estimating the level of human lethal violence (Supplementary Information section 9b). In addition, because the level of violence varies among populations of the same species<sup>10,20,21</sup>, all models include intraspecific variation in the level of mammalian lethal violence. The phylogenetically inferred level of lethal violence, averaging

across all models, was  $2.0 \pm 0.02\%$  of all deaths (Fig. 3a). These estimates seem to be robust to many potential biases, such as phylogenetic uncertainty, phylogenetic depth, sampling effort, and phylogeny size (Supplementary Information sections 3–6). Territoriality and sociability affect the phylogenetic inference of the level of lethal violence, as it was  $1.9 \pm 0.01\%$  in the models without these two variables but  $2.1 \pm 0.02\%$





**Figure 2 | Social behaviour and territoriality influence lethal aggression in mammals.** The figure shows the phylogenetically corrected level of lethal aggression per group (mean  $\pm$  s.e.m.) and the number of mammalian species included in each group. We used a phylogenetic generalized linear model (PGLS) to test the effect of territoriality (yes or no) and social behaviour (social or solitary) on lethal aggression. The level of lethal aggression was more intense in social and territorial species (PGLS,  $P < 0.05$  in all cases and mammal phylogenies; Extended Data Table 1), with no interaction between these two terms (Extended Data Table 1).

in the models including them (Fig. 3a). This is a consequence of *H. sapiens* being both social and territorial, two characteristics associated with a stronger tendency towards lethal violence in mammals (Fig. 2).

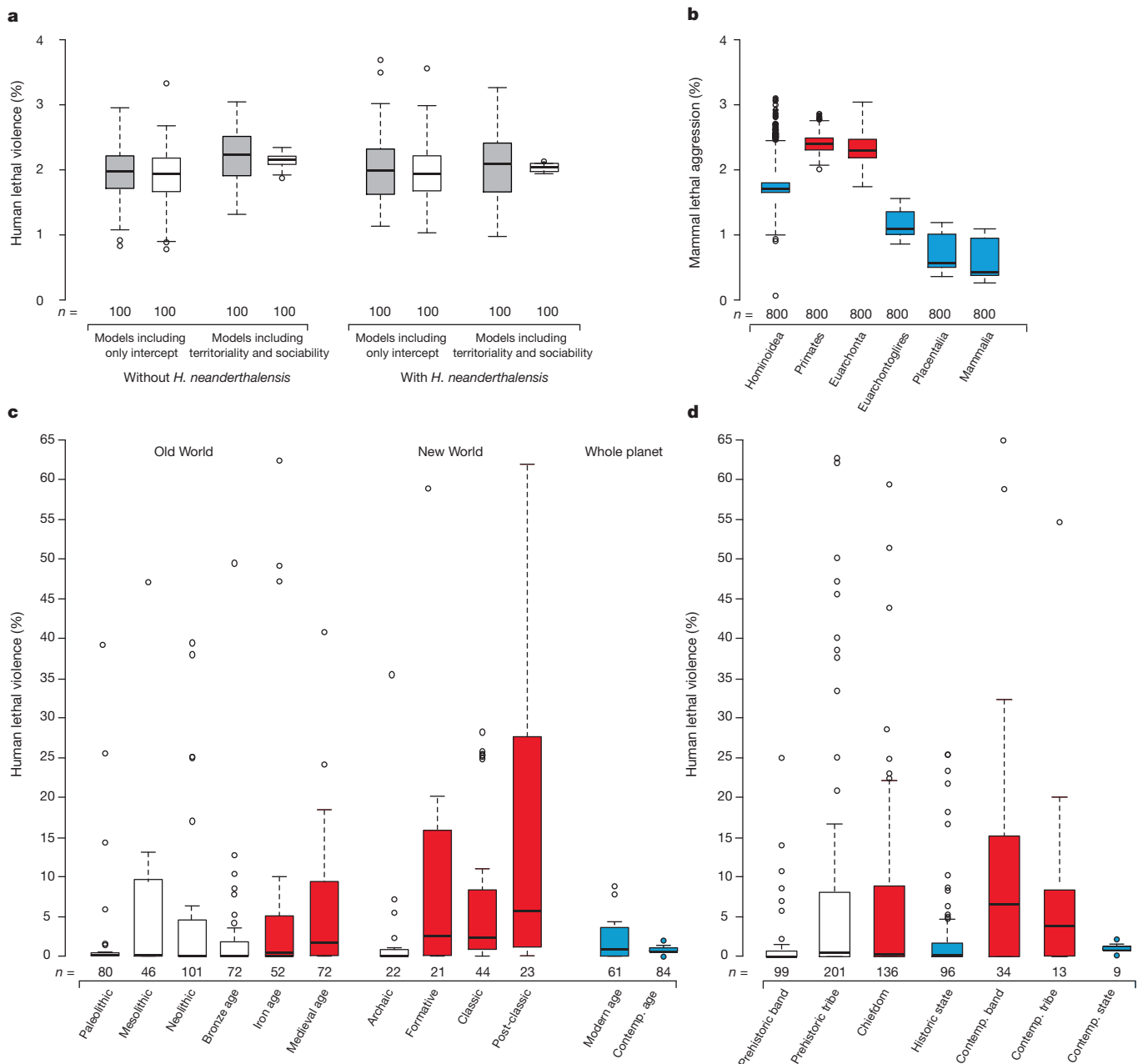
We subsequently explored how the level of lethal violence has changed during our evolutionary history by comparing it with the phylogenetically inferred level of lethal violence in relevant ancestral nodes that describe the course of human evolution (Fig. 1). The level of lethal violence was low in the most basal nodes, increasing to  $2.3 \pm 0.1\%$  of all deaths in the two nodes closely related with the origin of primates and slightly decreasing to  $1.8 \pm 0.1\%$  of all deaths in the ancestral ape (Fig. 3b). These results suggest that lethal violence is deeply rooted in the primate lineage.

We then compared whether the phylogenetically inferred level of lethal violence differed from the level empirically observed in human populations. The samples were categorized according to their age, using the standard periods from the New and Old World chronologies<sup>27</sup>. These data must be interpreted cautiously, because there was extensive intra-period variation in lethal violence. Nevertheless, a clear temporal pattern emerged (Fig. 3c). The level of lethal violence during human prehistory did not differ from the phylogenetic predictions (Fig. 3c). This result contrasts with some previous observations<sup>9,11</sup>, probably

because we have included more populations in our study and weighted all the analyses by the number of individuals per sample. The level of lethal violence during most historic periods was higher than the phylogenetic predictions for both humans (Fig. 3c and Supplementary Information section 7) and the ancestral Hominoidea (Fig. 3b). However, on entering the Modern and Contemporary ages (defined in Methods), the level of lethal violence decreased markedly, as previously reported<sup>11</sup> (Fig. 3c). Several potential biases may affect these results. The level of lethal violence inferred from skeletal remains could be underestimated because many deadly injuries do not damage the bones<sup>8</sup>. Nevertheless, no underestimation was detected for the periods in which both skeletal remains and statistical yearbooks are available (Supplementary Information section 7). Similarly, the presence of battlefields may artificially overestimate the level of lethal violence. However, the periods with highest level of lethal violence were not those with more organized intergroup conflicts (Supplementary Information section 8). Thus, the temporal pattern in the level of lethal violence seems to hold even after considering these potential biases. Concomitant changes in the cultural and ecological human environment may have caused this pattern. Notably, population density, a common ecological driver of lethal aggression in mammals<sup>18,21</sup>, was lower in periods with high levels of lethal violence than in the less violent Modern and Contemporary ages. High population density is therefore probably a consequence of successful pacification, rather than a cause of strife<sup>7</sup>.

Socio-political organization is a factor widely invoked to explain changes in violence<sup>5,7,11</sup>. To assess this effect, we classified human populations into four types<sup>28</sup>: bands, tribes, chiefdoms and states. Levels of lethal violence in prehistoric bands and tribes did not differ from the phylogenetic inferences (Fig. 3d). However, lethal violence is common in present-day bands and tribes (Fig. 3d), possibly because there are more detailed data on mortality from living people than from archaeological records. Nevertheless, some authors suggest that the level of lethal violence has increased in hunter-gatherers because they now live in denser populations in which intergroup conflicts are more likely<sup>3</sup>, or because they have contacted colonial societies where warfare or interpersonal violence is frequent<sup>29</sup>. The level of lethal violence in chiefdoms was also higher than the phylogenetic inferences (Fig. 3d). Severe violence has been frequently reported in chiefdoms<sup>30</sup>, mostly caused by territorial disputes, population and resource pressures, and competition for political status<sup>30</sup>. Finally, the level of lethal violence in state societies was lower than the phylogenetic inferences (Fig. 3d). It is widely acknowledged that monopolization of the legitimate use of violence by the state significantly decreases violence in state societies<sup>11,30</sup>.

In this study, we have explored the origin and evolution of human lethal violence by integrating a phylogenetic approach with an empirical analysis of lethal violence in human populations. The phylogenetic analysis suggests that a certain level of lethal violence in humans arises from the occupation of a position within a particularly violent mammalian clade, in which violence seems to have been ancestrally present. This means that humans have phylogenetically inherited their propensity for violence. We believe that this phylogenetic effect entails more than a mere genetic inclination to violence. In fact, social behaviour and territoriality, two behavioural traits shared with relatives of *H. sapiens*, seem to have also contributed to the level of lethal violence phylogenetically inherited in humans. Our analysis of human lethal violence shows that lethal violence in prehistoric humans matches the level inferred by our phylogenetic analyses, suggesting that we were, at the dawn of humankind, as violent as expected considering the common mammalian evolutionary history. This prehistoric level of lethal violence has not remained invariant but has changed as our history has progressed, mostly associated with changes in the socio-political organization of human populations. This suggests that culture can modulate the phylogenetically inherited lethal violence in humans.



**Figure 3 | Lethal violence in humans. a–d**, Box plots showing **a**, the phylogenetic inferences of human lethal violence assessed as the percentage of human deaths caused by conspecifics. These estimates were achieved through phylogenetic generalized linear models and correspond to the ancestral node of the tree rooted at the node separating *H. sapiens* from the rest of the mammals. All models were performed after logit-transforming the dependent variable and considering the intraspecific variation in mammal lethal aggression. Phylogenetic uncertainty was incorporated by using the tree provided by Fritz *et al.*<sup>22</sup> (grey colour) and a set of 100 randomly sampled trees from Faurby and Svenning<sup>23</sup> (white colour). **b**, The lethal aggression inferred for six important ancestral nodes of human evolution (apes, primates, Euarchonta, Euarchontoglires, placental mammals, and all mammals). **c**, Human lethal violence during

different temporal periods of human history, according to the Old World and New World chronologies<sup>27</sup>. **d**, Human lethal violence in different socio-political organizations<sup>28</sup>. In all cases the boxplots show median values, 50th percentile values (box outline), 95th percentile values (whiskers), and outlier values (circles). We tested whether the level of lethal violence observed in each ancestral node, human period and human socio-political organization differed significantly from the phylogenetic inferences in **a**. Colour indicates whether the observed lethal violence was statistically similar (white), higher (red), or lower (blue) than the phylogenetic inferences (Extended Data Tables 2, 3). In **a** and **b**, *n* indicates the number of iterations and in **c** and **d** it indicates the number of human populations (see Supplementary Information sections 7, 9c for the number of deaths).

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 March; accepted 15 August 2016.

Published online 28 September 2016.

1. Kelly, R. C. The evolution of lethal intergroup violence. *Proc. Natl Acad. Sci. USA* **102**, 15294–15298 (2005).

- Archer, J. The nature of human aggression. *Int. J. Law Psychiatry* **32**, 202–208 (2009).
- Bowles, S. Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science* **324**, 1293–1298 (2009).
- Wrangham, R. W. & Glowacki, L. Intergroup aggression in chimpanzees and war in nomadic hunter-gatherers: evaluating the chimpanzee model. *Hum. Nat.* **23**, 5–29 (2012).
- Fry, D. P. & Söderberg, P. Lethal aggression in mobile forager bands and implications for the origins of war. *Science* **341**, 270–273 (2013).

6. Sussman, R. W. in *War, Peace, and Human Nature: the Convergence of Evolutionary and Cultural Views* (ed. Fry, D. P.) 97–111 (Oxford Univ. Press, 2013).
7. Morris, I. *War! What is it Good For? Conflict and the Progress of Civilization from Primates to Robots* (Farrar, Straus & Giroux, 2014).
8. Martin, D. L. & Harrod, R. P. Bioarchaeological contributions to the study of violence. *Am. J. Phys. Anthropol.* **156**, (Suppl. 59), 116–145 (2015).
9. Keeley, L. H. *War Before Civilization* (Oxford Univ. Press, 1996).
10. Wrangham, R. & Peterson, D. *Demonic Males: Apes and the Origin of Human Violence* (Mariner Books, 1996).
11. Pinker, S. *The Better Angels of our Nature* (Viking Press, 2011).
12. Ferguson, R. B. in *War, Peace, and Human Nature: the Convergence of Evolutionary and Cultural Views* (ed. Fry, D. P.) 191–240 (Oxford Univ. Press, 2013).
13. Anholt, R. R. H. & Mackay, T. F. C. Genetics of aggression. *Annu. Rev. Genet.* **46**, 145–164 (2012).
14. Huber, R. & Brennan, P. A. Aggression. *Adv. Genet.* **75**, 1–6 (2011).
15. Daly, M. & Wilson, M. *Homicide* (Aldine de Gruyter, 1988).
16. Low, B. S. *Why Sex Matters: a Darwinian Look at Human Behavior* (Princeton Univ. Press, 2010).
17. Packer, C. & Pusey, A. E. in *Infanticide, Comparative and Evolutionary Perspectives* (eds Hausfater, G. & Hrdy, S. B.) 31–42 (Aldine Transactions, 1984).
18. Cubaynes, S. *et al.* Density-dependent intraspecific aggression regulates survival in northern Yellowstone wolves (*Canis lupus*). *J. Anim. Ecol.* **83**, 1344–1356 (2014).
19. Polis, G. A., Myers, C. A. & Hess, W. R. A survey of intraspecific predation within the class Mammalia. *Mammal Rev.* **14**, 187–198 (1984).
20. Lukas, D. & Huchard, E. Sexual conflict. The evolution of infanticide by males in mammalian societies. *Science* **346**, 841–844 (2014).
21. Archer, J. *The Behavioural Biology of Aggression* (Cambridge Univ. Press, 1984).
22. Fritz, S. A., Bininda-Emonds, O. R. & Purvis, A. Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics. *Ecol. Lett.* **12**, 538–549 (2009).
23. Faurby, S. & Svenning, J. C. A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach. *Mol. Phylogenet. Evol.* **84**, 14–26 (2015).
24. Opie, C., Atkinson, Q. D., Dunbar, R. I. & Shultz, S. Male infanticide leads to social monogamy in primates. *Proc. Natl Acad. Sci. USA* **110**, 13328–13332 (2013).
25. Garland, T. Jr & Ives, A. R. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat.* **155**, 346–364 (2000).
26. Goberna, M. & Verdú, M. Predicting microbial traits with phylogenies. *ISME J.* **10**, 959–967 (2016).
27. Shaw, I. & Jameson, R. A *Dictionary of Archaeology* (Blackwell, 1999).
28. Johnson, A. W. & Earle, T. K. *The Evolution of Human Societies: From Foraging Group to Agrarian State* (Stanford Univ. Press, 2000).
29. Allen, M. W. & Jones, T. L. *Violence and Warfare Among Hunter-Gatherers* (Left Coast Press, 2014).
30. Abrutyn, S. & Lawrence, K. From chiefdom to state: toward an integrative theory of the evolution of polity. *Sociol. Perspect.* **53**, 419–442 (2010).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The authors thank E. W. Schupp, P. Jordano, M. Lineham, J. A. Carrión, M. Goberna, A. Montesinos, J. G. Martínez, C. Sánchez Prieto, R. Torices, R. Menéndez and F. Perfectti for comments on an early version of this manuscript.

**Author Contributions** The study was conceived by J.M.G. Data were compiled by all authors. Analysis was performed by M.V., J.M.G. and A.G.M. All authors discussed the results and contributed to the manuscript.

**Author Information** The data used in this study are available in Supplementary Information section 9. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M.G. ([jmgreyes@eeza.csic.es](mailto:jmgreyes@eeza.csic.es)).

**Reviewer Information** *Nature* thanks O. Bininda-Emonds, M. Pagel and M. L. Wilson for their contribution to the peer review of this work.



## METHODS

No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

**Lethal aggression in mammals.** To estimate lethal aggression in mammals (defined as the percentage of deaths caused by conspecifics) we compiled a database including the amount of conspecific killing observed in many species of mammals. We conducted computer searches including the words (alone or in combination): 'mammal', 'mortality factors', 'causes of mortality', 'infanticide', 'death', 'conspecific mortality', 'conspecific fighting', 'intraspecific aggression' and 'conspecific aggression', as well as some other words related to relevant mortality factors in some mammal species, such as 'bushmeat', 'road killing' and 'overhunting'. We pooled all sources of conspecific mortality (active and passive infanticide, intergroup aggression, cannibalism and intraspecific predation, male–male fighting during mating period, territorial defensive behaviour, maternal abandonment, accidental injury). We considered only lethal conspecific interactions, ignoring non-lethal aggression, because the recording of aggressive interactions ending in the death of any of the interacting organisms, both in humans and non-human mammals, is more precise<sup>8</sup>. We found information about more than four million deaths in the 1,024 mammal species (~20% of the total species) from 137 families (~80% of total families) and the three main extant mammalian clades (Prototheria, Metatheria and Eutheria) (Supplementary Information section 9a). We obtained information from several studies in order to incorporate the intraspecific variability in lethal aggression for each mammal species. For each mammal included in our database, we recorded its territoriality (yes or no) and social behaviour (social or solitary) using information compiled in the Animal Diversity Web (<http://www.animaldiversity.org>).

**Mammal phylogeny.** The phylogenetic relationship between the mammals included in the database was built using Fritz *et al.*<sup>22</sup> and Faurby and Svenning<sup>23</sup> phylogenies, which are updated phylogenies of the supertree of Bininda-Emonds *et al.*<sup>31</sup>, to account for the more recent mammalian taxonomy of Wilson and Reeder<sup>32</sup>. First, we used the phylogeny provided by Fritz *et al.*<sup>22</sup> including 5,020 extant mammals. Afterwards, we used a set of 100 phylogenies provided by Faurby and Svenning<sup>23</sup> that contains 5,747 extant and extinct mammals (including species with dated records from the Late Pleistocene, defined as the last 130,000 years). Using this set of phylogenies, we were able to incorporate phylogenetic uncertainty in all our analyses. In each phylogeny we pruned all species not included in the database and, in the few cases in which a species was missing in the supertree, we selected the closest relative (usually, a congeneric species, see Supplementary Information section 9a). Mortality data about subspecies were pooled at the species level.

We performed additional analyses with the inclusion of *H. neanderthalensis* because: i) close relatives of modern humans can be very informative to estimate their phylogenetically shared traits, and ii) including fossils in the phylogeny results in more reliable ancestral state reconstructions<sup>33</sup>. The Faurby and Svenning<sup>22</sup> phylogeny includes *H. neanderthalensis*. However, the Fritz *et al.*<sup>23</sup> phylogeny only contains extant species. For this reason, we grafted *H. neanderthalensis* into this latter phylogeny, indicating an evolutionary divergence from *H. sapiens* 0.43 million years ago (Mya)<sup>34</sup> and extinction 0.028 Mya<sup>35</sup>. Although these dates are contested<sup>36</sup>, variations of a few thousand years did not significantly alter the phylogenetic prediction of human lethal violence. For example, when time of divergence was changed to 0.23 Mya, the mean prediction remained the same but with a slightly higher confidence interval. The level of lethal violence in *H. neanderthalensis* was obtained from multiple sources (see Supplementary Information section 9b).

**Lethal violence in humans.** To estimate lethal violence in humans (defined as the percentage of people that died owing to interpersonal violence) we compiled information from almost 600 human populations and societies spanning from the Palaeolithic to the present (Supplementary Information section 9c). Because of the extremely wide temporal range, we obtained information derived from very disparate sources, namely bioarchaeological and palaeo-osteological reports, ethnographic records, statistical yearbooks and verbal autopsies (a method to determine probable causes of death when no medical record or formal medical attention is available; they are performed by non-medical field workers, recording written narratives from reliable informants in local languages that describe the events that preceded the death). Owing to this heterogeneity, and because our goal was to compare the level of lethal violence in humans with the level of lethal aggression in mammals, we did not differentiate the specific causes of intraspecific mortality. Rather, we pooled together the deaths caused by war, homicide, manslaughter, infanticide, sacrifice, cannibalism and so on, without differentiating whether lethal events involved only one perpetrator or were coalitional and collective killings. Although it is worth investigating how specific types of violence have evolved in humans, we could not explore this issue because some types of violence have been insufficiently studied, both in non-human mammals (for example, inter-group aggression in social mammals other than chimpanzees) and humans (for example, infanticide in historical

societies). Lethal violence was determined for each source using the criteria of the researchers. Ethnographic records, statistical yearbooks and verbal autopsies commonly included the casualties of the interpersonal violence. The death toll owing to interpersonal violence in bioarchaeological studies was found by following the most widely used criterion in this type of study; that is, the presence of perimortem and blade injuries as an indication of death caused by interpersonal violence<sup>8,37</sup>. This means that we did not include antemortem and healed injuries in our calculation of lethal interpersonal violence<sup>37</sup>. Nevertheless, skeletal trauma should be viewed as minimal estimates, since many injuries caused by conspecifics do not damage the bones<sup>8,38</sup>.

The samples were categorized according to their age and socio-political organization. To assign the age to each sample, we considered the periods used to divide human history according to both the New World and Old World chronologies<sup>27</sup>. Old World human societies were grouped into Paleolithic (~50,000–12,000 BP), Mesolithic (~12,000–10,200 BP), Neolithic/Calcolithic (~10,200–5,000 BP), Bronze Age (~5,300–3,200 BP), Iron Age (~3,200–1,300 BP) and Medieval periods (~1,300–500 BP). New World human societies were grouped in Archaic (~12,000–3,000 BP), Formative (~3,000–1,500 BP), Classic (~1,500–800 BP) and Post-Classic periods (~800–500 BP). From then on, we considered two further periods affecting human societies throughout the entire world, the Modern Age (~500–100 BP) and the Contemporary Age (~100 BP–present day).

We followed the widely accepted socio-political classification<sup>28,39</sup>, according to which human societies can be classified into four types: bands (small, nomadic, egalitarian groups of people, usually hunter–gatherers), tribes (small, mostly egalitarian, groups with limited social rank usually resident in permanent villages as hunter–horticulturalists), chiefdoms (stratified, hierarchical non-industrial societies usually based on kinship) and states (politically organized complex societies). To assign each sample to different socio-political and temporal categories, we relied on the information from each original source (Supplementary Information section 8c). The use of standard statistics to summarize information coming from disparate sources with extremely different sample sizes and time coverage is problematic, as has been reported<sup>40</sup>. To avoid such issues, we pooled all the samples (skeletal remains, dead individuals and so on) found during each period (see Supplementary Information section 8c for an exhaustive list of cases, samples and studies) and depicted them using box plots.

**Phylogenetic signal of mammal lethal aggression.** The phylogenetic signal for lethal aggression was calculated using Pagel's lambda<sup>41</sup> that compares the similarity of the covariances among species with the covariances expected under Brownian evolution. Significant phylogenetic signal occurs when  $\lambda > 0$  and may take values of either  $0 < \lambda < 1$  (indicating that close relatives resemble each other less than expected under Brownian evolution) or  $\lambda = 1$  (indicating that close relatives are as similar as would be expected under Brownian motion). Values of  $\lambda > 1$  (indicating that close relatives are more similar than expected by Brownian evolution) cannot be reached because the off-diagonal elements in the variance–covariance matrix cannot be larger than the diagonal elements<sup>42</sup>. To account for the possibility of a phylogenetic signal higher than expected under Brownian motion, we also calculated Blomberg's *K* (that is, the ratio between the observed phylogenetic signal and that expected under a Brownian evolution model)<sup>43</sup>. This phylogenetic signal metric is not restricted in its upper limit, and ranges from 0 (no phylogenetic signal) to infinity, with  $K = 1$  indicating Brownian evolution. Statistical significance of Pagel's  $\lambda$  was calculated through a likelihood ratio test, comparing the likelihood of the model that was fitted to the data to that of a model in which  $\lambda$  was fixed to 0. Significance of Blomberg's *K* was calculated through a randomization test from a null model constructed with 1,000 random permutations of the data across the tips of the mammal tree. Both tests were performed using the R package 'phytools'<sup>44</sup>. The level of phylogenetic signal of lethal aggression in mammals measured as Blomberg's *K* ( $K = 0.09$ ) was significantly higher than 0 ( $P = 0.013$ ) and lower than 1 ( $P \ll 0.001$ ). This indicates that close relatives tend to have similar values of lethal violence but at a level lower than would be expected under Brownian evolution. This evolutionary pattern is consistent with that shown by Pagel's lambda ( $\lambda = 0.60$ ) and therefore only this metric is shown in the main text. The evolution of lethal aggression throughout the phylogeny of mammals was estimated using stochastic mapping as implemented in the R package 'phytools'<sup>44</sup>. Lethal aggression was logit-transformed before all analyses.

**Effect of territoriality and sociability on mammal lethal aggression.** To examine which factors explained the level of lethal aggression in mammals, we performed a phylogenetic generalized-least-squares (PGLS) model<sup>45</sup>, with lethal aggression (logit-transformed) as the dependent variable and territoriality and sociability as independent variables. PGLS takes into account the phylogenetic signal in the residuals of the model fitted to the data<sup>45</sup>. To account for the intraspecific variability in lethal aggression, for each of the 1,024 mammal species, we generated a normal distribution of lethal aggression values with their empirically observed means and standard errors. To control for potential biases produced by between-study

differences in sample size, the means and standard errors that were used to generate the random distributions were first weighted by the number of individuals included in each study. We then ran the analysis 100 times, randomly sampling each time a value from each of the 1,024 normal distributions. When a species was represented by a single value, we used as its standard error the across-species average of standard errors. The analyses were run with the help of the PGLS command in the R package 'caper'<sup>46</sup>.

**Phylogenetic estimation of human lethal violence.** Phylogenetic trait estimation techniques were used to obtain the lethal violence level for *H. sapiens* as a function of its position in the mammal phylogeny. These techniques take advantage of ancestral state estimation methods to predict traits of extant species<sup>25,47</sup>. The trait value of the focal species can be estimated as the ancestral node of the tree rerooted at the most recent common ancestor of the focal species and the rest of the tree<sup>48,49</sup>. The trait value estimated with this ancestral estimation method is the same as that provided by the intercept of a PGLS performed on the same tree. However, PGLS allows us to simultaneously include the level of the phylogenetic signal and other traits as covariates to improve the phylogenetic estimation of the study trait<sup>25</sup>. Following this approach, we also estimated human lethal violence with the help of a PGLS approach with territoriality and sociability as covariates and the phylogenetic information of the mammal tree rooted in the node where *H. sapiens* diverged from the rest of the mammals. The target species must be excluded from the analysis to estimate the PGLS parameters. Four PGLS models were fitted to our data: (i) without covariates and without *H. neanderthalensis*; (ii) with territoriality and sociability as factorial covariates but without *H. neanderthalensis*; (iii) without covariates and with *H. neanderthalensis*; and (iv) with territoriality and sociability as factorial covariates and with *H. neanderthalensis*. In all models, the dependent variable was logit-transformed and its variance was included using the approach explained in the previous section.

**Lethal aggression in main ancestral nodes of the human lineage.** We estimated levels of lethal aggression in the most recent common ancestor of six important clades defining the course of the evolutionary history of humans: the class Mammalia, the infraclass Placentalia (placental mammals), the superorder Euarchontoglires or Supraprimates (primates, tree-shrews, colugos, rodents and hares), the grandorder Euarchonta (primates, colugos and tree-shrews), the order Primates (primates) and the superfamily Hominoidea (apes). Lethal aggression in these ancestral nodes was inferred using the same analytical approach as that used to estimate lethal violence in humans.

**Accuracy of the estimation of mammal lethal aggression from the PGLS.** The accuracy of trait-estimation in a particular species increases with the level of phylogenetic signal of the study trait<sup>25</sup>. To test for the accuracy of our models under the observed phylogenetic signal, we used leave-one-out cross-validations with the whole mammalian data set in Supplementary Information section 9a. We inferred the level of lethal violence (logit-transformed) for each mammal species with the PGLS procedure and compared it with its actual value. We first examined the relationship between the estimated and observed lethal violence values<sup>50</sup> and subsequently calculated the proportion of species for which the actual value fell inside the 95% confidence interval of the estimated trait (Supplementary Information section 2).

**Effect of sampling effort on the estimation of human lethal violence.** To check whether the estimates of conspecific-mediated human mortality were influenced by inappropriate or insufficient sampling, we repeated all analyses considering the subset of mammalian species with more than 50 observations ( $n = 645$  mammals). We performed PGLS analysis to test whether territorial and social behaviour still influence the level of lethal aggression (logit-transformed) for this subset of well-sampled species. Afterwards, we calculated the conspecific-mediated human mortality using this subset of well-sampled mammals (Supplementary Information section 4).

**Effect of phylogenetic depth on the estimation of human lethal violence.** To check whether the estimates of conspecific-mediated human mortality were influenced by the depth of the phylogeny, we repeated these analyses by progressively including deeper nodes to obtain the estimate and the 95% confidence intervals using the PGLS model without covariates. We considered the following hierarchically nested clades, from shallower to deeper: Homininae, Hominidae, Hominoidea, Catarrhini, Simiiformes, Haplorrhini, Primates, Primatomorpha, Euarchonta, Euarchontoglires, Boreoeutheria, Eutheria, Theriiformes and Mammalia<sup>51</sup>. We are aware that moving from shallower to deeper nodes means including an increasing number of species in the analyses (for example, we have only four Homininae but 1,022 Theriiformes in our phylogeny). To subsequently check whether the increasing number of species has any effect on the 95% confidence intervals, we repeated all analyses with random-pruned phylogenies equalling the number of species included in each of the clades described here (50 random phylogenies per clade) (Supplementary Information section 5).

**Effect of phylogeny size on the estimation of human lethal violence.** To check whether the estimates of conspecific-mediated human mortality were influenced

by the size of the phylogeny, we repeated these analyses with the progressive inclusion of more species in the phylogenies. Specifically, we estimated human lethal violence and its 95% confidence interval in 50 randomly generated phylogenies with 100, 200, 300, 400, 500, 800, 900 and 1,000 spp., using the PGLS model without covariates. Afterwards, we contrasted these values with the level of human lethal violence obtained using the empirical phylogeny, checking whether smaller phylogenies departed from empirical results more strongly than larger phylogenies (Supplementary Information section 6).

**Statistical difference between phylogenetically estimated lethal violence in humans and ancestral nodes.** We have checked whether the level of lethal violence phylogenetically inferred in humans is different from the lethal aggression inferred for the main ancestral nodes using *t*-tests. The phylogenetic estimates of both lethal violence in humans and lethal aggression in ancestral mammals were obtained by joining the 100 values obtained for each of the four PGLS models (with and without covariates and with and without *H. neanderthalensis*) and the two mammalian phylogenies used (Fritz *et al.*<sup>22</sup> and Faurby and Svenning<sup>23</sup> phylogenies). We subsequently tested, by means of *t*-tests, whether these two distributions differed. Because we repeated the same test six times (once per ancestral node), we corrected all *P* values by means of sequential Bonferroni corrections.

**Statistical difference between observed and phylogenetically estimated lethal violence.** For each temporal period and socio-political organization, we randomly sampled a given value of observed mortalities from a normal distribution with the same mean and standard error and compared it with a randomly sampled, phylogenetically estimated value. The phylogenetically estimated values were obtained by joining the 100 values obtained for each of the four PGLS models (with and without covariates and with and without *H. neanderthalensis*) and the two mammalian phylogenies (Fritz *et al.*<sup>22</sup> and Faurby and Svenning<sup>23</sup> phylogenies). We repeated these paired comparisons 800 times, and recorded the proportion of times where the observed values were higher or lower than the phylogenetically estimated values. We subsequently tested, by means of binomial tests, whether this proportion differed from the randomly expected deviation. We ran each binomial test 1,000 times and retained the average *P* values and deviance from the expected value. All *P* values shown underwent sequential Bonferroni correction.

- Bininda-Emonds, O. R. P. *et al.* The delayed rise of present-day mammals. *Nature* **446**, 507–512 (2007); Corrigendum **456**, 274 (2008).
- Wilson, D. E. & Reeder, D. M. *Mammal Species of the World: a Taxonomic and Geographic Reference*, 2nd–3rd edn. (Smithsonian Institution Press / John Hopkins Univ. Press, 1993–2005).
- Finarelli, J. A. & Flynn, J. J. Ancestral state reconstruction of body size in the Caniformia (Carnivora, Mammalia): the effects of incorporating data from the fossil record. *Syst. Biol.* **55**, 301–313 (2006).
- Finlayson, C. *et al.* Late survival of Neanderthals at the southernmost extreme of Europe. *Nature* **443**, 850–853 (2006).
- Arsuaga, J. L. *et al.* Neandertal roots: Cranial and chronological evidence from Sima de los Huesos. *Science* **344**, 1358–1363 (2014).
- Hublin, J. J. The origin of Neandertals. *Proc. Natl Acad. Sci. USA* **106**, 16022–16027 (2009).
- Mays, S. *The Archaeology of Human Bones* (Routledge, 2010).
- Milner, G. R. Nineteenth-century arrow wounds and perceptions of prehistoric warfare. *Am. Antiq.* **70**, 144–156 (2005).
- Service, E. R. *Profiles in Ethnology* (Harpercollins College Div., 1963).
- War, peace, and human nature: the Convergence of Evolutionary and Cultural Views (ed. Fry, D. P.) (Oxford Univ. Press, 2013).
- Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
- Münkemüller, T. *et al.* How to measure and test phylogenetic signal. *Methods Ecol. Evol.* **3**, 743–756 (2012).
- Blomberg, S. P., Garland, T. Jr & Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**, 717–745 (2003).
- Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).
- Freckleton, R. P., Harvey, P. H. & Pagel, M. Phylogenetic analysis and comparative data: a test and review of evidence. *Am. Nat.* **160**, 712–726 (2002).
- Orme, A. D. *et al.* caper: Comparative analyses of phylogenetics and evolution in R (v.0.5.2). <https://cran.r-project.org/web/packages/caper/index.html> (2013).
- Martins, E. P. & Hansen, T. F. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *Am. Nat.* **149**, 646–667 (1997).
- Kembel, S. W., Wu, M., Eisen, J. A. & Green, J. L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLOS Comput. Biol.* **8**, e1002743 (2012).
- Nunn, C. & Zhu, L. in *Modern Phylogenetic Comparative Methods and their Application in Evolutionary Biology* (ed. Garamszegi, L. Z.) 481–514 (Springer, 2014).
- Piñeiro, G., Perelman, S., Guerschman, J. P. & Paruelo, J. M. How to evaluate models: observed vs. predicted or predicted vs. observed? *Ecol. Modell.* **216**, 316–322 (2008).
- Brand, S. J. *Systema Naturae 2000. The Taxonomicon* (Amsterdam, 2005).

**Extended Data Table 1 | Outcome of the phylogenetic generalized linear model testing the effect of territoriality and social behaviour on the magnitude of lethal aggression in mammal species ( $n = 1,024$  species)**

	Estimate $\pm$ s.e.m	t-value	p-value
<b>Fritz et al.'s phylogeny</b>			
Territoriality	0.54 $\pm$ 0.50	3.80	0.001
Social behaviour	0.47 $\pm$ 0.51	2.71	0.014
Territoriality * Social behaviour	0.48 $\pm$ 0.55	1.33	0.244
lambda of the model	0.54		0.0001
<b>Faurby &amp; Svenning's phylogeny</b>			
Territoriality	0.53 $\pm$ 0.51	3.01	0.001
Social behaviour	0.47 $\pm$ 0.51	2.70	0.009
Territoriality * Social behaviour	0.48 $\pm$ 0.51	0.43	0.341
lambda of the model	0.88		0.0001

We performed this analysis using the mammalian phylogeny provided by Fritz *et al.*<sup>22</sup> and 100 mammalian phylogenies provided by Faurby and Svenning<sup>23</sup>. In this latter case, we show the across-phylogeny mean of each statistical parameter. Lethal aggression was logit-transformed before all analyses.



**Extended Data Table 2 | Outcome of the *t*-tests assessing difference between the inferred value of lethal violence at each of the chosen ancestral nodes in the mammalian phylogeny and the phylogenetic estimates of human lethal violence**

Ancestral Nodes	t-test	p-value	Significance
Class Mammalia (mammals)	- 72.49	0.0001	YES
Infraclass Placentalia (placentals)	- 70.88	0.0001	YES
Superorder Euarchontoglires (primates, rodents, hares)	- 50.50	0.0001	YES
Grandorder Euarchonta (primates, tree-shrews, colugos)	15.66	0.0001	YES
Order Primates (primates)	20.78	0.0001	YES
Superfamily Hominoidea (apes)	- 16.31	0.0001	YES

We compared the lethal aggression of the ancestral nodes with the magnitudes of lethal violence obtained according to the four PGLS models (with and without covariates and with and without *H. neanderthalensis*) and the two mammalian phylogenies (Fritz *et al.*<sup>22</sup> and Faurby and Svenning<sup>23</sup> phylogenies) using a *t*-test. Significance after sequential Bonferroni correction at  $\alpha = 0.05$ .

**Extended Data Table 3 | Outcome of the binomial tests assessing difference between the observed lethal violence in human societies and the inferred lethal violence according to the phylogenetic analysis**

	<b>Difference between the observed and the phylogenetically inferred lethal violence</b>	<b>p-value</b>	<b>Significance</b>
<b>Temporal periods</b>			
<b>Old World Chronology</b>			
Paleolithic	+ 0.0 %	0.522	NO
Mesolithic	+ 4.6 %	0.064	NO
Neolithic	+ 5.4 %	0.029	NO
Bronze Age	+ 2.5 %	0.272	NO
Iron Age	+ 8.1 %	0.002	YES
Medieval Age	+ 7.7 %	0.002	YES
<b>New World Chronology</b>			
Archaic	+ 4.5 %	0.073	NO
Formative	+ 6.2 %	0.0001	YES
Classic	+ 16.5 %	0.0001	YES
Postclassic	+ 13.0 %	0.0001	YES
<b>The Entire World</b>			
Modern Age	- 37.1 %	0.0001	YES
Contemporary Age	- 23.7 %	0.0001	YES
<b>Type of society</b>			
Historic Band	+ 4.6 %	0.131	NO
Historic Tribe	+ 4.4 %	0.116	NO
Historic Chiefdom	+ 5.7 %	0.009	YES
Historic State	- 42.9 %	0.0001	YES
Contemporary Band	+ 18.5 %	0.00001	YES
Contemporary Tribe	+ 12.2 %	0.003	YES
Contemporary State	- 27.4 %	0.00001	YES

We compared the observed lethal violence of each type of human society with the magnitudes of lethal violence obtained according to the four PGLS models (with and without covariates and with and without *H. neanderthalensis*) and the two mammalian phylogenies (Fritz *et al.*<sup>22</sup> and Faurby and Svenning<sup>23</sup> phylogenies). Each binomial test was run 1000 times. Significance after sequential Bonferroni correction at  $\alpha = 0.05$ .

# Genomic analyses inform on migration events during the peopling of Eurasia

A list of authors and affiliations appears at the end of the paper

High-coverage whole-genome sequence studies have so far focused on a limited number<sup>1</sup> of geographically restricted populations<sup>2–5</sup>, or been targeted at specific diseases, such as cancer<sup>6</sup>. Nevertheless, the availability of high-resolution genomic data has led to the development of new methodologies for inferring population history<sup>7–9</sup> and refuelled the debate on the mutation rate in humans<sup>10</sup>. Here we present the Estonian Biocentre Human Genome Diversity Panel (EGDP), a dataset of 483 high-coverage human genomes from 148 populations worldwide, including 379 new genomes from 125 populations, which we group into diversity and selection sets. We analyse this dataset to refine estimates of continent-wide patterns of heterozygosity, long- and short-distance gene flow, archaic admixture, and changes in effective population size through time as well as for signals of positive or balancing selection. We find a genetic signature in present-day Papuans that suggests that at least 2% of their genome originates from an early and largely extinct expansion of anatomically modern humans (AMHs) out of Africa. Together with evidence from the western Asian fossil record<sup>11</sup>, and admixture between AMHs and Neanderthals predating the main Eurasian expansion<sup>12</sup>, our results contribute to the mounting evidence for the presence of AMHs out of Africa earlier than 75,000 years ago.

The paths taken by AMHs out of Africa (OoA) have been the subject of considerable debate over the past two decades. Fossil and archaeological evidence<sup>13,14</sup>, and craniometric studies<sup>15</sup> of African and Asian populations, demonstrate that *Homo sapiens* was present outside of Africa ~120–70 thousand years ago (kya)<sup>11</sup>. However, this colonization has been viewed as a failed expansion OoA<sup>16</sup> since genetic analyses of living populations have been consistent with a single OoA followed by serial founder events<sup>17</sup>.

Ancient DNA (aDNA) sequencing studies have found support for admixture between early Eurasians and at least two archaic human lineages<sup>18,19</sup>, and suggest modern humans reached Eurasia at around 100 kya<sup>12</sup>. In addition, aDNA from modern humans suggests population structuring and turnover, but little additional archaic admixture, in Eurasia over the last 35–45 thousand years<sup>20–22</sup>. Overall, these findings indicate that the majority of human genetic diversity outside Africa derives from a single dispersal event that was followed by admixture with archaic humans<sup>18,23</sup>.

We used ADMIXTURE to analyse the genetic structure in our diversity set (Extended Data Figs 1, 2; Supplementary Information 1.1–7). We further compared the individual-level haplotype similarity of our samples using fineSTRUCTURE (Extended Data Fig. 3). Despite small sample sizes, we inferred 106 genetically distinct populations forming 12 major regional clusters, corresponding well to the 148 self-identified population labels. This clustering forms the basis for the groupings used in the scans of natural selection. Similar genetic affinities are highlighted by plotting the outgroup  $f_3$  statistic<sup>9</sup> in the form  $f_3(X, Y; \text{Yoruba})$ , which here measures shared drift between a non-African population  $X$  and any modern or ancient population  $Y$  from Yoruba as an African outgroup (Supplementary Information 2.2.6, Extended Data Fig. 4).

Our sampling allowed us to consider geographic features correlated with gene flow by spatially interpolating genetic similarity measures

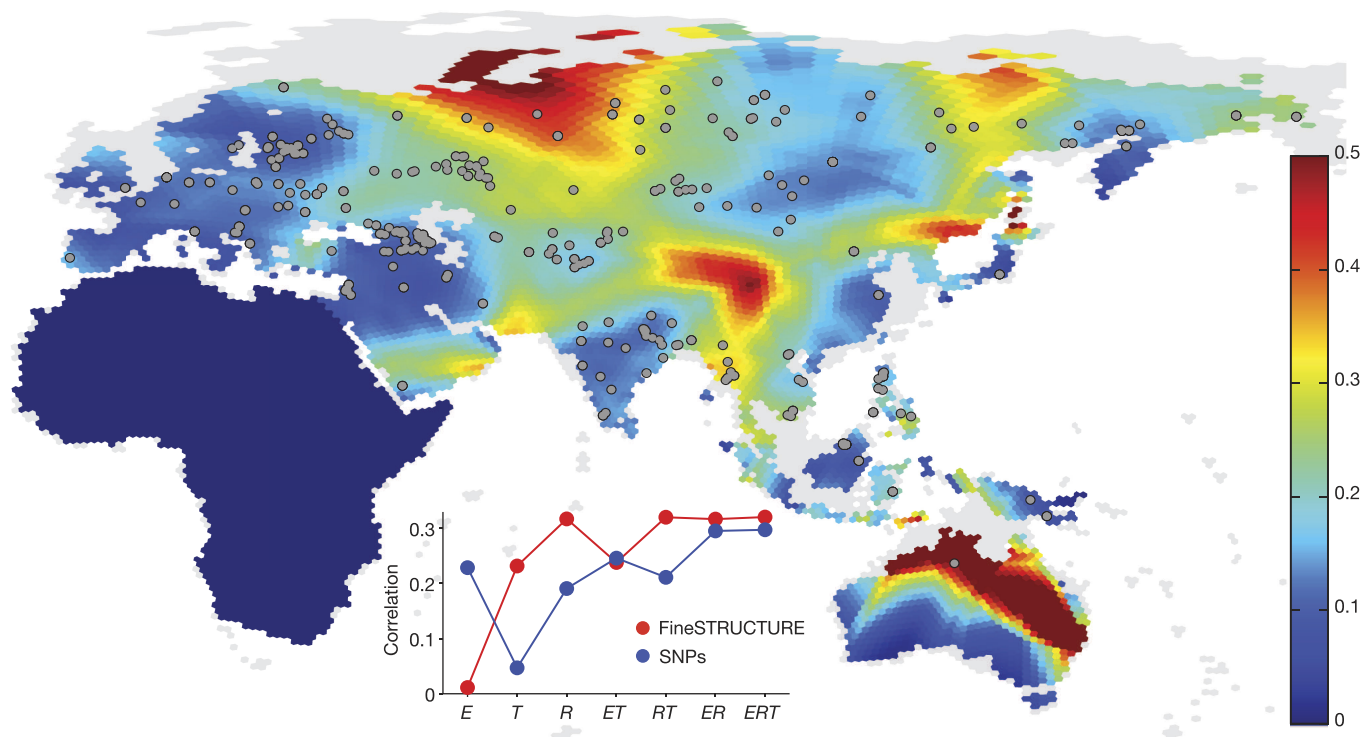
between pairs of populations (Supplementary Information 2.2.2). We considered several measures and report gradients of allele frequencies in Fig. 1, which was compared to gene flow patterns from EEMS<sup>24</sup> as a validation (Extended Data Fig. 5). Controlling for pairwise geographic distance, we find a correlation between these genetic gradients and geographic and climatic features such as precipitation and elevation (inset of Fig. 1, Supplementary Information 2.2.2).

We screened for evidence of selection by first focusing on loci that showed the highest allelic differentiation among groups (Supplementary Information 3). We then performed positive and purifying selection scans (Methods), and found some candidate loci that replicate previously known and functionally supported findings (Supplementary Table 1:3.3.4-I, Supplementary Information 3.1, Extended Data Fig. 6; Supplementary Table 1:3.1-IV, VI). Additionally, we infer more purifying selection in Africans in genes involved in pigmentation (bootstrapping  $p$  value (bpv) for  $R_{X/Y}$  scores  $< 0.05$ ) (Extended Data Fig. 6) and immune response against viruses (bpv  $< 0.05$ ), while further purifying selection was indicated on olfactory receptor genes in Asians (bpv  $< 0.05$ ) (Supplementary Table 1:3.1.1-II). Our scans for ancient balancing selection found a significant enrichment (FDR  $< 0.01$ ) of antigen processing/presentation, antigen binding, and MHC and membrane component genes (Supplementary Information 3.2 and 3.3, Supplementary Table 1:3.3.2-I–III). The HLA (*HLA-C*)-associated gene (*BTNL2*) was the top highest scoring candidate in 8 of 12 geographic regions for the HKA test (Supplementary Table 1:3.3.1-I). Our positive selection scans, variant-based analyses (Supplementary Information 3.2 and 3.3) and gene enrichment studies also suggest new candidate loci (Supplementary Information 3.4 and 3.5, Supplementary Table 1:3.5-I–VI), a subset of which is highlighted in Supplementary Table 1:3-I.

Using fineSTRUCTURE, we find in the genomes of Papuans and Philippine Negritos more short haplotypes assigned as African than seen in genomes for individuals from other non-African populations (Extended Data Fig. 7). This pattern remains after correcting for potential confounders such as phasing errors and sampling bias (Supplementary Information 2.2.1). These shorter shared haplotypes would be consistent with an older population split<sup>25</sup>. Indeed, the Papuan–Yoruban median genetic split time (using multiple sequential Markovian coalescent (MSMC)) of 90 kya predates the split of all mainland Eurasian populations from Yorubans at ~75 kya (Supplementary Table 1:2.2.3-I, Extended Data Fig. 4, Fig. 2a). This result is robust to phasing artefacts (Extended Data Fig. 8, see Methods). Furthermore, the Papuan–Eurasian MSMC split time of ~40 kya is only slightly older than splits between west Eurasian and East Asian populations dated at ~30 kya (Extended Data Fig. 4). The Papuan split times from Yoruba and Eurasia are therefore incompatible with a simple bifurcating population tree model.

At least two main models could explain our estimates of older divergence dates for Sahul populations from Africa than mainland Eurasians in our sample: 1) admixture in Sahul with a potentially un-sampled archaic human population that split from modern humans either before or at the same time as did Denisova and Neanderthal; or 2) admixture in Sahul with a modern human population (extinct OoA line; xOoA) that left Africa after the split between modern humans





**Figure 1 | Genetic barriers across space.** Spatial visualization of genetic barriers inferred from genome-wide genetic distances, quantified as the magnitude of the gradient of spatially interpolated allele frequencies (value denoted by colour bar; grey areas have been land during the last glacial maximum but are currently underwater). Here we used a spatial kernel smoothing method based on the matrix of pairwise average heterozygosity and a MATLAB script that plots the hexagons of the grid with a colour coding to represent gradients. Inset, partial correlation

between magnitude of genetic gradients and combinations of different geographic factors, elevation (*E*), temperature (*T*) and precipitation (*R*), for genetic gradients from fineSTRUCTURE (red) and allele frequencies (blue). This analysis (Supplementary Information 2.2.2 for details) shows that genetic differences within this region display some correlation with physical barriers such as mountain ranges, deserts, forests, and open water (such as the Wallace line).

and Neanderthals, but before the main expansion of modern humans in Eurasia (main OoA).

We consider support for these two non-mutually exclusive scenarios. Because the introgressing lineage has not been observed with aDNA, standard methods are limited in their ability to distinguish between these hypotheses. Furthermore, we show (Supplementary Information 2.2.7) that single-site statistics, such as Patterson's  $D^{9,18}$  and sharing of non-African Alleles (nAAs), are inherently affected by confounding effects owing to archaic introgression in non-African populations<sup>23</sup>. Our approach therefore relies on multiple lines of evidence using haplotype-based MSMC and fineSTRUCTURE comparisons (which we show should have power at this timescale<sup>26</sup>; Supplementary Information 2.2.13).

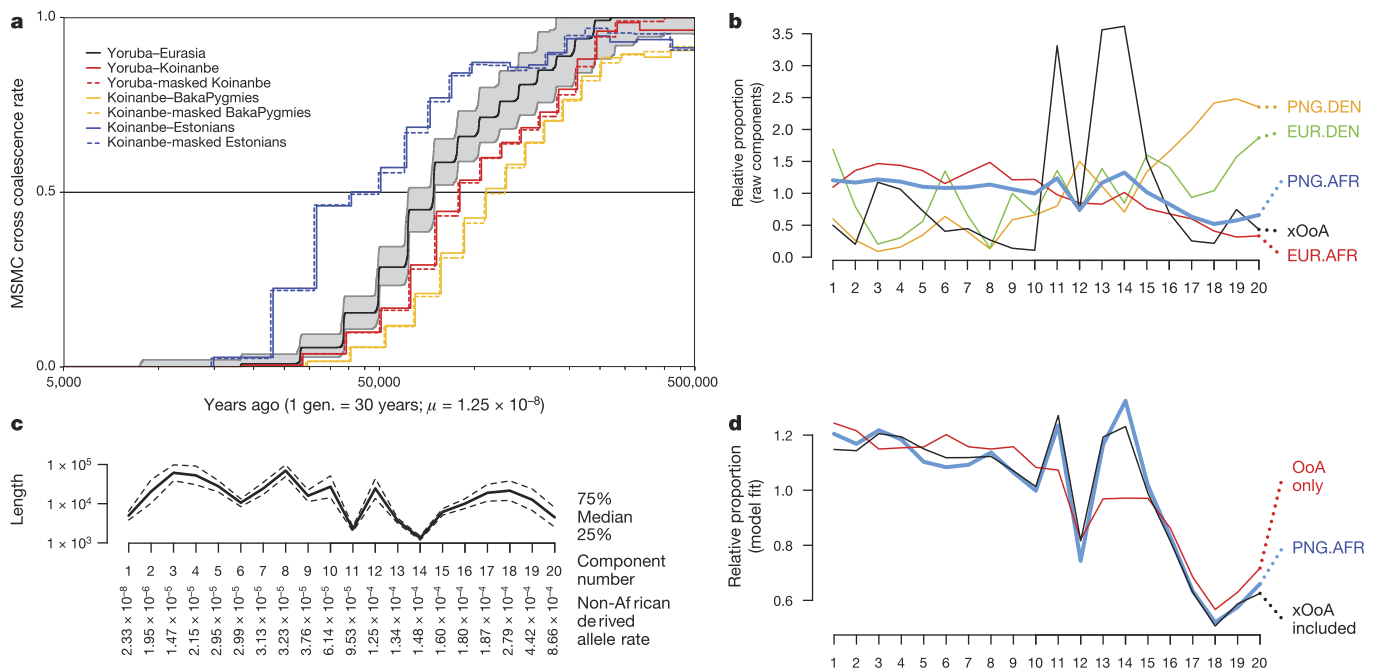
We located and masked putatively introgressed<sup>27</sup> Denisova haplotypes from the genomes of Papuans, and evaluated phasing errors by symmetrically phasing Papuans and Eurasians genomes (Methods). Neither modification (Fig. 2a, Supplementary Information 2.2.9, Supplementary Table 1:2.2.9-I) changed the estimated split time (based on MSMC) between Africans and Papuans (Methods, Supplementary Information 2.2.8, Extended Data Fig. 8, Supplementary Table 1:2.2.8-I). MSMC dates behave approximately linearly under admixture (Extended Data Fig. 8), implying that the hypothesized lineage may have split from most Africans around 120 kya (Supplementary Information 2.2.4 and 2.2.8).

We compared the effect on the MSMC split times of an xOoA or a Denisova lineage in Papuans by extensive coalescent simulations (Supplementary Information 2.2.8). We could not simulate the large Papuan–African and Papuan–Eurasian split times inferred from the data, unless assuming an implausibly large contribution from a Denisova-like population. Furthermore, while the observed shift in the African–Papuan MSMC split curve can be qualitatively reproduced

when including a 4% genomic component that diverged 120 kya from the main human lineage within Papuans, a similar quantity of Denisova admixture does not produce any significant effect (Extended Data Fig. 8). This favours a small presence of xOoA lineages rather than Denisova admixture alone as the likely cause of the observed deep African–Papuan split. We also show (Methods) that such a scenario is compatible with the observed mitochondrial DNA and Y chromosome lineages in Oceania, as also previously argued<sup>13,28</sup>.

We further tested our hypothesized xOoA model by analysing haplotypes in the genomes of Papuans that show African ancestry not found in other Eurasian populations. We re-ran fineSTRUCTURE adding the Denisova, Altai Neanderthal and the Human Ancestral Genome sequences<sup>29</sup> to a subset of the diversity set. FineSTRUCTURE infers haplotypes that have a most recent common ancestor (MRCA) with another individual. Papuan haplotypes assigned as African had, regardless, an elevated level of non-African derived alleles (that is, nAAs fixed ancestral in Africans) compared to such haplotypes in Eurasians. They therefore have an older mean coalescence time with our African samples.

Owing to the deep divergence between the sampled Denisova and the one introgressed into modern humans, it is possible that some archaic haplotypes have a MRCA with an African instead of Denisova and are assigned as 'African'. We can resolve the coalescence time, and hence origin, of these haplotypes by their sequence similarity with modern Africans. To account for the archaic introgression we modelled these genomic segments as a mixture of haplotypes assigned a) as African or b) as Denisova in Eurasians and c) haplotypes assigned as Denisova in Papuans. These haplotypes are modelled (see Methods, Extended Data Fig. 9) in terms of the distribution of length and mutation rate measured as a density of non-African derived alleles. Since Eurasians (specifically Europeans) have not experienced Denisova admixture,



**Figure 2 | Evidence of an xOoA signature in the genomes of modern Papuans. a**, MSMC split times plot. The Yoruba-Eurasia split curve shows the mean of all Eurasian genomes against one Yoruba genome. The grey area represents top and bottom 5% of runs. We chose a Koinanbe genome as representative of the Sahul populations. **b–d**, Decomposition of Papuan haplotypes inferred as African by fineSTRUCTURE. **b**, Semi-parametric decomposition of the joint distribution of haplotype lengths and non-African derived allele rate per SNP, showing the relative proportion of haplotypes in  $K=20$  components of the distribution, ordered by non-African derived allele rate, relative to the overall proportion of

this approach disentangles lineages that coalesce before the human/Denisova split from those that coalesce after.

We found that the xOoA signature (Fig. 2b–d; Supplementary Information 2.2.10) was necessary to account for the number of short haplotypes with ‘moderate’ nAAs density in the data (that is, proportion of non-African-derived sites higher than that of Eurasian haplotypes assigned as African but significantly lower than that of those assigned Denisova in either Eurasians or Papuans). Consistent with our MSMC findings (Supplementary Information 2.2.4), xOoA haplotypes have an estimated MRCA 1.5 times older than the Eurasian haplotypes in Papuan genomes, while the Denisovan haplotypes in Papuans are four times older than the Eurasian haplotypes. Adding up the contributions across the genome (Methods) leads to a genome-wide estimate of 1.9% xOoA (95% confidence interval 1.5–3.3) in Papuans, which we view as a lower bound.

Our results consistently point towards a contribution from a modern human source for derived<sup>29</sup> alleles that are found in the genome sequence of Papuans but not in Africans. Possible confounders could involve a shorter generation time in Papuan and Philippine Negrito populations<sup>30</sup>, different recombination processes, or alternative demographic histories that have not been investigated here. We therefore strongly encourage the development of new model-based approaches that can investigate further the haplotype patterns described here.

In conclusion, our results suggest that while the genomes of modern Papuans derive primarily from the main expansion of modern humans out of Africa, we estimate that at least 2% of their genome sequence reflects an earlier, otherwise extinct, dispersal (Extended Data Fig. 10).

The inferred date of the xOoA split time ( $\sim 120$  kya) is consistent with fossil and archaeological evidence for an early expansion of *H. sapiens* from Africa<sup>13,14</sup>. Furthermore, the recently identified modern human admixture into the Altai Neanderthal before 100 kya<sup>12</sup> is consistent with a modern human presence outside Africa well

before the main OoA split time ( $\sim 75$  kya). Further studies will confirm whether the Papuan genetic signature reported here and the one observed in Altai Neanderthals reflect the same xOoA human group, as well as clarify the timing and route followed during such an early expansion. The high similarity between Papuans and the Altai Neanderthal reported in Extended Data Fig. 1 may indeed reflect a shared xOoA component. Further studies are needed to explore this model and suggest that understanding human evolutionary history will require the recovery of aDNA from additional fossils, and further archaeological investigations in under-explored geographical regions.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 20 September 2015; accepted 24 August 2016.**

**Published online 21 September 2016.**

- Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
- Lachance, J. *et al.* Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).
- Pagani, L. *et al.* Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96**, 986–991 (2015).
- Clemente, F. J. *et al.* A selective sweep on a deleterious mutation in CPT1A in Arctic populations. *Am. J. Hum. Genet.* **95**, 584–589 (2014).
- Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
- Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
- Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
- Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).

9. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
10. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13**, 745–753 (2012).
11. Grove, M. *et al.* Climatic variability, plasticity, and dispersal: a case study from Lake Tana, Ethiopia. *J. Hum. Evol.* **87**, 32–47 (2015).
12. Kuhlwillm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016).
13. Groucutt, H. S. *et al.* Rethinking the dispersal of *Homo sapiens* out of Africa. *Evol. Anthropol.* **24**, 149–164 (2015).
14. Liu, W. *et al.* The earliest unequivocally modern humans in southern China. *Nature* **526**, 696–699 (2015).
15. Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc. Natl Acad. Sci. USA* **111**, 7248–7253 (2014).
16. Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc. Natl Acad. Sci. USA* **110**, 10699–10704 (2013).
17. Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15**, R159–R160 (2005).
18. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
19. Reich, D. *et al.* Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am. J. Hum. Genet.* **89**, 516–528 (2011).
20. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
21. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23**, 553–559 (2013).
22. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
23. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
24. Petkova, D., Novembre, J. & Stephens, M. Visualizing spatial population structure with estimated effective migration surfaces. *Nat. Genet.* **48**, 94–100 (2016).
25. Hellenthal, G. *et al.* A genetic atlas of human admixture history. *Science* **343**, 747–751 (2014).
26. Chapman, N. H. & Thompson, E. A. A model for the length of tracts of identity by descent in finite random mating populations. *Theor. Popul. Biol.* **64**, 141–150 (2003).
27. Wall, J. D. *et al.* Higher levels of Neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209 (2013).
28. Posth, C. *et al.* Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. *Curr. Biol.* **26**, 827–833 (2016).
29. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
30. Migliano, A. B., Vinicius, L. & Lahr, M. M. Life history trade-offs explain the evolution of human pygmies. *Proc. Natl Acad. Sci. USA* **104**, 20216–20219 (2007).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** Support was provided by: Estonian Research Infrastructure Roadmap grant no 3.2.0304.11-0312; Australian Research Council Discovery grants (DP110102635 and DP140101405) (D.M.L., M.W. and E.W.); Danish National Research Foundation; the Lundbeck Foundation and KU2016 (E.W.); ERC Starting Investigator grant (FP7 - 261213) (T.K.); Estonian Research Council grant PUT766 (G.C. and M.K.); EU European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre (R.V.; M.Me. and A.Me.), and Centre of Excellence for Genomics and Translational Medicine Project No. 2014-2020.4.01.15-0012 to EGC of UT (A.Me.) and EBC (M.Me.); Estonian Institutional Research grant IUT24-1 (L.S., M.J., A.K., B.Y., K.T., C.B.M., Le.S., H.Sa., S.L., D.M.B., E.M., R.V., G.H., M.K., G.C., T.K. and M.Me.) and IUT20-60 (A.Me.); French Ministry of Foreign and European Affairs and French ANR grant number ANR-14-CE31-0013-01 (F.-X.R.); Gates Cambridge Trust Funding (E.J.); ICG SB RAS (No. VI.58.1.1) (D.V.L.); Leverhulme Programme grant no. RP2011-R-045 (A.B.M., P.G. and M.G.T.); Ministry of Education and Science of Russia; Project 6.656.2014/K (S.A.F.); NEFEX grant funded by the European Union (People Marie Curie Actions; International Research Staff Exchange Scheme; call FP7-PEOPLE-2012-IRSES-number 318979) (M.Me., G.H. and M.K.); NIH grants 5DP1ES022577 05, 1R01DK104339-01, and 1R01GM113657-01 (S.Tis.); Russian Foundation for Basic Research (grant N 14-06-00180a) (M.G.); Russian Foundation for Basic Research; grant 16-04-00890 (O.B. and E.B.); Russian Science Foundation grant 14-14-00827 (O.B.); The Russian Foundation for Basic Research (14-04-00725-a), The Russian Humanitarian Scientific Foundation (13-11-02014) and the Program of the Basic Research of the RAS Presidium “Biological diversity” (E.K.K.); Wellcome Trust and Royal Society grant WT104125AIA & the Bristol Advanced Computing Research Centre (<http://www.bris.ac.uk/acrc/>) (D.J.L.); Wellcome Trust grant 098051 (Q.A.; C.T.-S. and Y.X.); Wellcome Trust Senior Research Fellowship grant 100719/Z/12/Z (M.G.T.); Young Explorers Grant from the National Geographic Society (8900-11) (C.A.E.); ERC Consolidator Grant 647787 ‘LocalAdaptat’ (A.Ma.); Program of the RAS Presidium “Basic research for the development of the Russian Arctic” (B.M.); Russian Foundation for Basic Research grant 16-06-00303 (E.B.); a Rutherford Fellowship (RDF-10-MAU-001) from the Royal Society of New Zealand (M.P.C.).

**Author Contributions** R.V., E.W., T.K. and M.Me. conceived the study. A.K., K.T., C.B.M., Le.S., E.P., G.A., C.M., M.W., D.L., G.Z., S.T., D.D., Z.S., G.N.N.S., K.M., J.L., L.D.D., M.G., P.N., I.E., L.A.T., O.U., F.-X.R., N.B., H.S., T.L., M.P.C., N.A.B., V.S., L.A., D.Pr., H.Sa., M.Mo., C.A.E., D.V.L., S.A., G.C., J.T.S.W., E.Mi., A.Ka., S.L., R.K., N.T., V.A., I.K., D.M., L.Y., D.M.B., E.B., A.Me., M.D., B.M., M.V., S.A.F., L.P.O., M.Mi., M.L., A.B.M., O.B., E.K.K., E.M., M.G.T. and E.W. conducted anthropological research and/or sample collection and management. J.L. and S.Ti. provided access to data. L.P., D.J.L., E.J., A.Mo., A.E., M.Mi., F.C., G.H., M.D., L.S., J.W., A.C., R.M., M.A.W.S., S.K., C.I., C.L.S., M.J., M.K., G.S.J., T.A., F.M.I., A.K., Q.A., C.T.-S., Y.X., B.Y., C.B.M., T.K. and M.Me. analysed data. L.P., D.J.L., E.J., A.Mo., L.S., M.K., K.T., C.B.M., Le.S., G.C., M.Mi., P.G., M.L., A.B.M., M.P., E.M., M.G.T., A.Ma., R.N., R.V., E.W., T.K. and M.Me. contributed to the interpretation of results. L.P., D.J.L., E.J., A.Mo., A.E., F.C., G.H., M.D., A.C., M.A.W.S., B.Y., J.L., S.Ti., M.Mi., P.G., M.L., A.B.M., M.P., M.G.T., A.Ma., R.N., R.V., E.W., T.K. and M.Me. wrote the manuscript.

**Additional Information** The newly sequenced genomes are part of the Estonian Biocentre Human Genome Diversity Panel (EGDP) and were deposited in the ENA archive under accession number PRJEB12437 and are also freely available through the Estonian Biocentre website ([www.ebc.ee/free\\_data](http://www.ebc.ee/free_data)). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.P. ([lp.lucapagani@gmail.com](mailto:lp.lucapagani@gmail.com)), T.K. ([tk331@cam.ac.uk](mailto:tk331@cam.ac.uk)) or M.Me. ([mait@ebc.ee](mailto:mait@ebc.ee)).

**Reviewer Information** Nature thanks R. Dennell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Luca Pagani<sup>1,2,3\*</sup>, Daniel John Lawson<sup>4\*</sup>, Evelyn Jagoda<sup>2,5\*</sup>, Alexander Mörseburg<sup>2\*</sup>, Anders Eriksson<sup>6,7\*</sup>, Mario Mitt<sup>8,9</sup>, Florian Clemente<sup>2,10</sup>, Georgi Hudjashov<sup>1,11,12</sup>, Michael DeGiorgio<sup>13</sup>, Lauri Saag<sup>1</sup>, Jeffrey D. Wall<sup>14</sup>, Alexia Cardona<sup>2,15</sup>, Reedik Mägi<sup>8</sup>, Melissa A. Wilson Sayres<sup>16,17</sup>, Sarah Kaewert<sup>2</sup>, Charlotte Inchley<sup>2</sup>, Christiana L. Scheib<sup>2</sup>, Mari Järve<sup>1</sup>, Monika Karmin<sup>1,11,18</sup>, Guy S. Jacobs<sup>19,20</sup>, Tiago Antao<sup>21</sup>, Florin Mircea Iliescu<sup>2</sup>, Alena Kushniarevich<sup>1,22</sup>, Qasim Ayub<sup>23</sup>, Chris Tyler-Smith<sup>23</sup>, Yali Xue<sup>23</sup>, Bayazit Yunusbayev<sup>1,24</sup>, Kristina Tambets<sup>1</sup>, Chandana Basu Mallick<sup>1</sup>, Lehti Saag<sup>18</sup>, Elvira Pocheshkhova<sup>25</sup>, George Andriadze<sup>26</sup>, Craig Muller<sup>27</sup>, Michael C. Westaway<sup>28</sup>, David M. Lambert<sup>28</sup>, Grigor Zoraqi<sup>29</sup>, Shahlo Turdikulova<sup>30</sup>, Dilbar Dalimova<sup>31</sup>, Zhaxyllyk Sabitov<sup>32</sup>, Gazi Nurun Nahar Sultana<sup>33</sup>, Joseph Lachance<sup>34,35</sup>, Sarah Tishkoff<sup>36</sup>, Kuvat Momynaliyev<sup>37</sup>, Jainagul Isakova<sup>38</sup>, Larisa D. Damba<sup>39</sup>, Marina Gubina<sup>39</sup>, Pagbajabyn Nymadawa<sup>40</sup>, Irina Evseeva<sup>41,42</sup>, Lubov Atramantova<sup>43</sup>, Olga Utevska<sup>43</sup>, Francois-Xavier Ricaut<sup>44</sup>, Nicolas Brucato<sup>44</sup>, Herawati Sudoyo<sup>45</sup>, Thierry Letellier<sup>44</sup>, Murray P. Cox<sup>12</sup>, Nikolay A. Barashkov<sup>46,47</sup>, Vedrana Škaro<sup>48,49</sup>, Lejla Mulahasanovic<sup>50</sup>, Dragan Primorac<sup>49,51,52,53</sup>, Hovhannes Sahakyan<sup>1,54</sup>, Maru Mormina<sup>55</sup>, Christina A. Eichstaedt<sup>5,56</sup>, Daria V. Lichman<sup>39,57</sup>, Syafiq Abdullah<sup>58</sup>, Gyaneshwer Chaubey<sup>1</sup>, Joseph T. S. Wee<sup>59</sup>, Evelin Mihailov<sup>6</sup>, Alexandra Karunas<sup>24,60</sup>, Sergei Litvinov<sup>1,24,60</sup>, Rita Khusainova<sup>24,60</sup>, Natalya Ekomasova<sup>60</sup>, Vita Akhmetova<sup>24</sup>, Irina Khidiyatova<sup>24,60</sup>, Damir Marjanovi<sup>61,62</sup>, Levon Yepiskoposyan<sup>54</sup>, Doron M. Behar<sup>1</sup>, Elena Balanovska<sup>63</sup>, Andres Metspalu<sup>8,9</sup>, Miroslava Derenko<sup>64</sup>, Boris Malyarchuk<sup>64</sup>, Mikhail Voevoda<sup>39,57,65</sup>, Sardana A. Fedorova<sup>46,47</sup>, Ludmila P. Osipova<sup>39,57</sup>, Marta Mirazon Lahr<sup>66</sup>, Pascale Gerbault<sup>67</sup>, Matthew Leavesley<sup>68,69</sup>, Andrea Bamberg Migliano<sup>70</sup>, Michael Petraglia<sup>71</sup>, Oleg Balanovsky<sup>63,72</sup>, Elza K. Khusnutdinova<sup>24,60</sup>, Ene Metspalu<sup>1,18</sup>, Mark G. Thomas<sup>67</sup>, Andrea Manica<sup>7</sup>, Rasmus Nielsen<sup>27,73</sup>, Richard Villems<sup>1,18,74\*</sup>, Eske Willerslev<sup>27\*</sup>, Toomas Kivisild<sup>1,2\*</sup> & Mait Metspalu<sup>1\*</sup>

<sup>1</sup>Estonian Biocentre, 51010 Tartu, Estonia. <sup>2</sup>Department of Archaeology and Anthropology, University of Cambridge, Cambridge CB2 1QH, UK. <sup>3</sup>Department of Biological, Geological and Environmental Sciences, University of Bologna, Via Selmi 3, 40126 Bologna, Italy. <sup>4</sup>Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK. <sup>5</sup>Department of Human Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA. <sup>6</sup>Integrative Systems Biology Lab, Division of Biological and Environmental Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia. <sup>7</sup>Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK. <sup>8</sup>Estonian Genome Center, University of Tartu, 51010 Tartu, Estonia. <sup>9</sup>Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia. <sup>10</sup>Institut de Biologie Computationnelle, Université Montpellier 2, 34095 Montpellier, France. <sup>11</sup>Department of Psychology, University of Auckland, Auckland 1142, New Zealand. <sup>12</sup>Statistics and Bioinformatics Group, Institute of Fundamental Sciences, Massey University, 4442 Palmerston North, New Zealand. <sup>13</sup>Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA. <sup>14</sup>Institute for Human Genetics, University of California, San Francisco, California 94143, USA. <sup>15</sup>MRC Epidemiology Unit, University of Cambridge, Institute of Metabolic Science, Box 285, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK. <sup>16</sup>School of Life Sciences, Arizona State University, Tempe, Arizona 85287, USA. <sup>17</sup>Center for Evolution and Medicine, The Biodesign Institute, Tempe, Arizona 85287, USA. <sup>18</sup>Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia. <sup>19</sup>Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK. <sup>20</sup>Institute for Complex Systems Simulation, University of Southampton, Southampton SO17 1BJ, UK. <sup>21</sup>Division of Biological Sciences,



University of Montana, Missoula, Montana 59812, USA. <sup>22</sup>Institute of Genetics and Cytology, National Academy of Sciences, BY-220072 Minsk, Belarus. <sup>23</sup>The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK. <sup>24</sup>Institute of Biochemistry and Genetics, Ufa Scientific Center of RAS, 450054 Ufa, Russia. <sup>25</sup>Kuban State Medical University, 350040 Krasnodar, Russia. <sup>26</sup>Scientific Research Center of the Caucasian Ethnic Groups, St. Andrews Georgian University, 0162 Tbilisi, Georgia. <sup>27</sup>Center for GeoGenetics, University of Copenhagen, 1350 Copenhagen, Denmark. <sup>28</sup>Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Nathan, Queensland 4111, Australia. <sup>29</sup>Center of Molecular Diagnosis and Genetic Research, University Hospital of Obstetrics and Gynecology, 1000 Tirana, Albania. <sup>30</sup>Center of High Technology, Academy of Sciences, 100047 Tashkent, Uzbekistan. <sup>31</sup>Institute of Bioorganic Chemistry Academy of Science, 100047 Tashkent, Uzbekistan. <sup>32</sup>L.N. Gumilyov Eurasian National University, 010008 Astana, Kazakhstan. <sup>33</sup>Centre for Advanced Research in Sciences (CARS), DNA Sequencing Research Laboratory, University of Dhaka, Dhaka-1000, Bangladesh. <sup>34</sup>Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6145, USA. <sup>35</sup>School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, USA. <sup>36</sup>Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6313, USA. <sup>37</sup>DNcode laboratories, 117623 Moscow, Russia. <sup>38</sup>Institute of Molecular Biology and Medicine, 720040 Bishkek, Kyrgyzstan. <sup>39</sup>Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 630090 Novosibirsk, Russia. <sup>40</sup>Mongolian Academy of Medical Sciences, 210620 Ulaanbaatar, Mongolia. <sup>41</sup>Northern State Medical University, 163000 Arkhangelsk, Russia. <sup>42</sup>Anthony Nolan, The Royal Free Hospital, Pond Street, London NW3 2QG, UK. <sup>43</sup>V. N. Karazin Kharkiv National University, 61022 Kharkiv, Ukraine. <sup>44</sup>Evolutionary Medicine group, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse, UMR 5288, Centre National de la Recherche Scientifique, Université de Toulouse 3, Toulouse 31073, France. <sup>45</sup>Genome Diversity and Diseases Laboratory, Eijkman Institute for Molecular Biology, 10430 Jakarta, Indonesia. <sup>46</sup>Department of Molecular Genetics, Yakut Scientific Centre of Complex Medical Problems, 677027 Yakutsk, Russia. <sup>47</sup>Laboratory of Molecular Biology, Institute of Natural Sciences, M.K. Ammosov North-Eastern Federal University, 677027 Yakutsk, Russia. <sup>48</sup>Genos DNA laboratory, 10000 Zagreb,

Croatia. <sup>49</sup>University of Osijek, Medical School, 31000 Osijek, Croatia. <sup>50</sup>Center for Genomics and Transcriptomics, CeGaT, GmbH, D-72076 Tübingen, Germany. <sup>51</sup>St. Catherine Specialty Hospital, 49210 Zabok and 10000 Zagreb, Croatia. <sup>52</sup>Eberly College of Science, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. <sup>53</sup>University of Split, Medical School, 21000 Split, Croatia. <sup>54</sup>Laboratory of Ethnogenomics, Institute of Molecular Biology, National Academy of Sciences, Republic of Armenia, 7 Hasratyan Street, 0014 Yerevan, Armenia. <sup>55</sup>Department of Applied Social Sciences, University of Winchester, Sparkford Road, Winchester SO22 4NR, UK. <sup>56</sup>Thoraxklinik Heidelberg, University Hospital Heidelberg, 69120 Heidelberg, Germany. <sup>57</sup>Novosibirsk State University, 630090 Novosibirsk, Russia. <sup>58</sup>RIPAS Hospital, Bandar Seri Begawan, BE1518 Brunei. <sup>59</sup>National Cancer Centre Singapore, 169610 Singapore. <sup>60</sup>Department of Genetics and Fundamental Medicine, Bashkir State University, 450000 Ufa, Russia. <sup>61</sup>Department of Genetics and Bioengineering, Faculty of Engineering and Information Technologies, International Burch University, 71000 Sarajevo, Bosnia and Herzegovina. <sup>62</sup>Institute for Anthropological Researches, 10000 Zagreb, Croatia. <sup>63</sup>Research Centre for Medical Genetics, Russian Academy of Sciences, Moscow 115478, Russia. <sup>64</sup>Genetics Laboratory, Institute of Biological Problems of the North, Russian Academy of Sciences, 685000 Magadan, Russia. <sup>65</sup>Institute of Internal Medicine, Siberian Branch of Russian Academy of Medical Sciences, 630009 Novosibirsk, Russia. <sup>66</sup>Leverhulme Centre for Human Evolutionary Studies, Department of Archaeology and Anthropology, University of Cambridge, Cambridge CB2 1QH, UK. <sup>67</sup>Research Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK. <sup>68</sup>Department of Archaeology, University of Papua New Guinea, University PO Box 320, 134 NCD, Papua New Guinea. <sup>69</sup>College of Arts, Society and Education, James Cook University, PO Box 6811, Cairns, Queensland 4870, Australia. <sup>70</sup>Department of Anthropology, University College London, London WC1H 0BW, UK. <sup>71</sup>Max Planck Institute for the Science of Human History, Kahlaische Strasse 10, D-07743 Jena, Germany. <sup>72</sup>Vavilov Institute for General Genetics, Russian Academy of Sciences, 119333 Moscow, Russia. <sup>73</sup>Department of Integrative Biology, University of California Berkeley, Berkeley 94720, California, USA. <sup>74</sup>Estonian Academy of Sciences, 6 Kohtu Street, Tallinn 10130, Estonia.

\*These authors contributed equally to this work.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Data preparation.** We analyse a set of genomes sequenced by the same technology (Complete Genomics Inc.) which results in minimal platform differences between batches of samples analysed by slight modifications of CG proprietary pipeline (Extended Data Fig. 2; Supplementary Information 1.6). Informed consent forms and REC approvals were obtained for all samples newly collected for this study. We see good concordance between CG sequence and Illumina genotyping array results for the same samples with minor reference bias in the latter data (Extended Data Fig. 2; Supplementary Information 1.6). In the final dataset, we retained only one second-degree (Australians, to make use of all the available samples) and five third-degree relatives pairs (Supplementary Table 1:1.7–I). All genomes were annotated against the Ensembl GRCh37 database and compared to dbSNP Human Build 141 and Phase 1 of the 1000 Genomes Project dataset<sup>29</sup> (Supplementary Information 1.1–1.6). We found 10,212,117 new SNPs, 401,911 of which were exonic. As expected from our sampling scheme, existing lists of variable sites have been extended mostly by the Siberian, Southeast Asian and South Asian genomes, which contribute 89,836 (22.4%), 63,964 (15.9%) and 40,758 (10.1%) of the new exonic variants detected in this study.

Compared to the genome-wide average, we see fewer heterozygous sites on chromosomes 1 and 2, and an excess on chromosomes 16, 19 and 21 (Extended Data Fig. 2). This pattern is independent of simple potential confounders, such as rough estimates of recombination activity and gene density (Supplementary Information 1.8), and mirrors the inter-chromosomal differences in divergence from chimpanzee<sup>31</sup>, suggesting large-scale differences in mutation rates among chromosomes. We confirmed this general pattern using 1000 Genomes Project data (Supplementary Information 1.8).

The ‘ancient genome diversity panel’ consisted of 106 samples from the main Diversity panel along with Altai Neanderthal, Denisova and the Modern Human reference genome. Sites that are heterozygous in archaic humans were removed.

**Geographic gradient analyses.** We used a Gaussian kernel smoothing (based on the shortest distance on land to each sample) to interpolate genetic patterns across space. Averaging over all markers, we obtained an expression for the mean square gradient of allele frequencies in terms of the matrix of genetic distance between pairs of samples (Supplementary Information 2.2.2). This provides a simple way to identify spatial regions that contribute strongly to genetic differences between samples, and can be used, in principle, for any measure of genetic difference (for fineSTRUCTURE data, we used negative shared haplotype length as a measure of differentiation).

To quantify the link between the magnitude of genetic gradients (from fineSTRUCTURE and allele frequency data) and geographic factors, we fitted a generalized linear model to the sum of genetic magnitude gradients on the shortest paths between samples to elevation, minimum quarterly temperature, and annual precipitation summed in the same way, controlling for path length and spatial random effects (Supplementary Information 2.2.2), and calculated partial correlations between genetic gradient magnitudes and geographic factors.

**FineSTRUCTURE analysis.** FineSTRUCTURE<sup>32</sup> was run as described in Supplementary Information 2.2.1. Within the 106 genetically distinct genetic groups, labels were typically genetically homogeneous—113 of the 148 population labels (76%) were assigned to only one ‘genetic cluster’. Similarly, genetic clusters were typically specific to a label, with 66 of the 106 ‘genetic clusters’ (62%) containing only one population label.

**Correction for phasing errors.** To check whether phasing errors could produce the shorter Papuan haplotypes, we focused on regions of the genome that had an extended (>500 kb) run of homozygosity. We ran ChromoPainter for each individual on only these regions, meaning each individual was only painted where it had been perfectly phased. This did not change the qualitative features (Supplementary Information 2.2.1).

**Removal of similar samples.** Papuans are genetically distinct from other populations due to tens of thousands of years of isolation. We wanted to check whether the length of haplotypes assigned as African was biased by the inclusion of a large number of relatively homogeneous Eurasians with few Papuans. To do this we repeated the  $n = 447$  painting allowing only donors from dissimilar populations, including only individuals who donated < 2% of a genome in the main painting. This did not change the qualitative haplotype length features (Supplementary Information 2.2.1).

**Inclusion of ancient samples.** We ran our smaller individual panel with ( $n = 109$ ) and without ( $n = 106$ ) ancient samples (Denisova, Neanderthal and ancestral human). This did not change the qualitative haplotype length features (Supplementary Information 2.2.1).

**Selection analyses.** We investigated balancing, positive and purifying selection for a part of the dataset with larger group sizes which was defined as the Selection subset (Supplementary Table 1:3.1–I and 3.2–I) using a wide range of window-based as well as variant-based approaches. Furthermore, we investigated how these signals relate to shared demographic history. Where possible we contextualized our findings by integrating them with information from various functional databases. Detailed descriptions of all methods used are available in Supplementary Information section 3.

**MSMC, Denisova masking, simulations of alternative scenarios and assessment of phasing robustness.** Genetic split times were initially calculated following the standard MSMC procedure<sup>8</sup>, and subsequently modified as follows. To estimate the effect of archaic admixture, putative Denisova haplotypes were identified in Papuans using a previously published method<sup>27</sup> and masked from all the analysed genomes. Particularly, whether a putative archaic haplotype was found in heterozygous or homozygous state within the chosen Papuan genome, the ‘affected’ locus was inserted into the MSMC mask files and, hence, removed from the analysis.

We note that a fraction of the Denisova and Neanderthal contributions to the Papuan genomes may be indistinguishable, owing to the shared evolutionary history of these two archaic populations. As a result, some of the removed ‘Denisova’ haplotypes may have actually entered the genome of Papuans through Neanderthal. Regardless of this, our exercise successfully shows that the MSMC split time estimates are not affected by the documented presence of archaic genomic component (whether coming entirely from Denisova or partially shared with Neanderthal).

We further excluded the role of Denisova admixture in explaining the deeper African–Papuan MSMC split times through coalescent simulations (using ms to generate 30 chromosomes of 5 Mbp each, and simulating each scenario 30 times). These showed that the addition of 4% Denisova lineages to the Papuan genomes does not change the MSMC results, while the addition of 4% xOoA lineages recreates the qualitative shift observed in the empirical data.

Phasing artefacts were also taken into account as putative confounders of the MSMC split time estimates. We re-ran MSMC after re-phasing one Estonian, one Papuan and 20 West African and Pygmy genomes in a single experiment. This way we ruled out potential artefacts stemming from the excess of Eurasian over Sahul samples during the phasing process. Both the archaic and phasing corrections yielded the same split time as of the standard MSMC runs.

**Emulation of all pairwise MSMC split times.** We confirmed that none of the other populations behaved as an outlier from those identified in the  $n = 22$  full pairwise analysis by estimating the MSMC split times between all pairs. We chose 9 representative populations (including Papuan, Yoruba and Baka) from the 22, and compared each of the 447 diversity panel genomes to them. For each individual  $l$  not in our panel, we obtain the positive mixture weights  $\alpha_k$  using the model

$$\hat{t}_{lj} = \sum_{k=1}^9 \alpha_{lk} t_{kj} \text{ for } j \in (1..9)$$

The parameters are estimated using the  $j \in (1..9)$  observations for which we have data using a quadratic loss function. We can then predict the unobserved values

$$\hat{t}_{li} = \sum_{k=1}^9 \alpha_{lk} t_{ki}$$

Examination of this matrix (Supplementary Information 2.2.3, Supplementary Table 1:2.2.3–III) implies no other populations are expected to have unusual MSMC split times from Africa.

**Mixture model for African haplotypes in Papuans.** *Obtaining haplotypes from painting.* We define African or Archaic haplotypes in Eurasians or Papuans as genomic loci spanning at least 1,000 bp, and showing SNPs that were assigned by chromopainter a  $\geq 50\%$  chance of copying from either an African or Archaic genome, respectively. For each haplotype we then calculated the number of non-African mutations, defined as sites found in derived state in a given haplotype and in ancestral state in all of the African genomes included in the present study. *Modelling.* We used a non-parametric model for the joint distribution of length and non-African derived allele mutation rate in haplotypes. We fit  $K = 20$  components to the joint distribution. Each component has a characteristic length  $l_k$ , variability  $\sigma_k$  and mutation rate  $\mu_k$ . A haplotype of length  $l_i$  with  $X_i$  such mutations from component  $I_i = k$  has the following distribution:

$$l_i | \{l_k, \sigma_k^2, I_i = k\} \sim \text{log-Normal}(l_k, \sigma_k^2)$$

$$X_i | \{l_k, \mu_k, I_i = k\} \sim \text{Binomial}(l_k, \mu_k)$$

This model for haplotype lengths is motivated by the extreme age of the split times we seek to model. Recent splits would lead to an exponential distribution of haplotype lengths. However, owing to haplotype fixation caused by finite population size, very old splits have finite (non-zero) haplotype lengths. Additionally, the data are left-censored since we cannot reliably detect haplotypes that are very short. We note that while this makes a single component a reasonable fit to the data, as  $K$  increases the specific choice becomes less important.

We then impose the prior  $p(I_i = k) = 1/K$  and use the expectation-maximization algorithm to estimate the mixture proportions  $\pi_{ik} = E(I_{ik}|I_i, X_i)$  along with the maximum likelihood parameter estimates  $\{l_k, \sigma_k^2, \mu_k\}$ . We do this for the four combinations of haplotypes assigned as African (AFR) and Denisova (DEN) found in Papuans (PNG) or Europeans (EUR), in order to learn the parameters. Supplementary Information 2.2.10 describes this in more detail. We then describe the distribution of haplotypes for each class  $c$  of haplotype in terms of the expected proportion of haplotypes found in each component,

$$\pi_{ck} = \frac{\pi'_{ck}}{\sum_{k=1}^K \pi'_{ck}} \text{ where } \pi'_{ck} = \sum_{i=1}^{N_c} \pi_{cik}$$

where  $N_c$  is the number of haplotypes of class  $c$ .  $\pi_c$  is a vector of the proportions from each of the  $K$  components.

**Single-out-of-Africa model.** We fit haplotypes assigned as African in Papuans as a mixture of the others in a second layer of mixture modelling:

$$\pi_{\text{PNG.AFR}} = \sum_{c \in \{\text{PNG.DEN, EUR.AFR, EUR.DEN}\}} \alpha_c \pi_c$$

where  $\alpha_c$  sum to 1. This is straightforward to fit.

**xOoA model.** We jointly estimate an additional component  $\pi_{\text{xOoA}}$  and the mixture contributions  $\beta_c$  under the mixture

$$\pi_{\text{PNG.AFR}} = \sum_{c \in \{\text{PNG.DEN, EUR.AFR, EUR.DEN, xOoA}\}} \beta_c \pi_c$$

This is non-trivial to fit. We use a penalization scheme to simultaneously ensure we a) obtain a valid mixture for  $\beta_c$ ; b) give a prediction  $x_k$  that is also a valid mixture; c) leave little signal in the residuals; and d) obtain a good fit. Cross-validation is used to obtain the optimal penalization parameters ( $A$  and  $B$ ) with the loss function:

$$\text{loss} = \sum_{k=1}^K e_k^2 + AP_A + BP_B,$$

where  $e_k$  are the residuals in each component,  $P_A = \left| \left( \sum_c \beta_c \right) - 1 \right| + \left| \left( \sum_k x_k \right) - 1 \right|$  (for a valid mixture) and  $P_B = s \cdot d(e_k)$  (for requirement c, good solutions will have

similar residuals across components). The loss is minimized via standard optimization techniques. Supplementary Information 2.2.10 details how initial values are found and explores the robustness of the solution to changes in  $A$  and  $B$ —the results do not change qualitatively for reasonable choices of these parameters, and the mixtures are valid to within numerical error.

**Genome-wide xOoA estimation.** We used the estimated xOoA derived allele mutation rate estimate  $\theta_{\text{xOoA}}$  to estimate the xOoA contribution in haplotypes classed as Eurasian or Papuan by ChromoPainter. First we obtained estimates of  $\pi_{\text{PNG.EUR}}$  and  $\pi_{\text{PNG.PNG}}$  using the single out-of-Africa model above, additionally allowing for a EUR.EUR contribution. We then estimate  $\alpha_{\text{xOoA}}$  using the observed mutation rate  $\theta_{\text{obs}}$  and that predicted under the mixture model  $\theta_{\text{mix}}$  by rearranging the mixture:

$$\theta_{\text{obs}} = \alpha_{\text{xOoA}} \theta_{\text{xOoA}} + (1 - \alpha_{\text{xOoA}}) \theta_{\text{mix}}$$

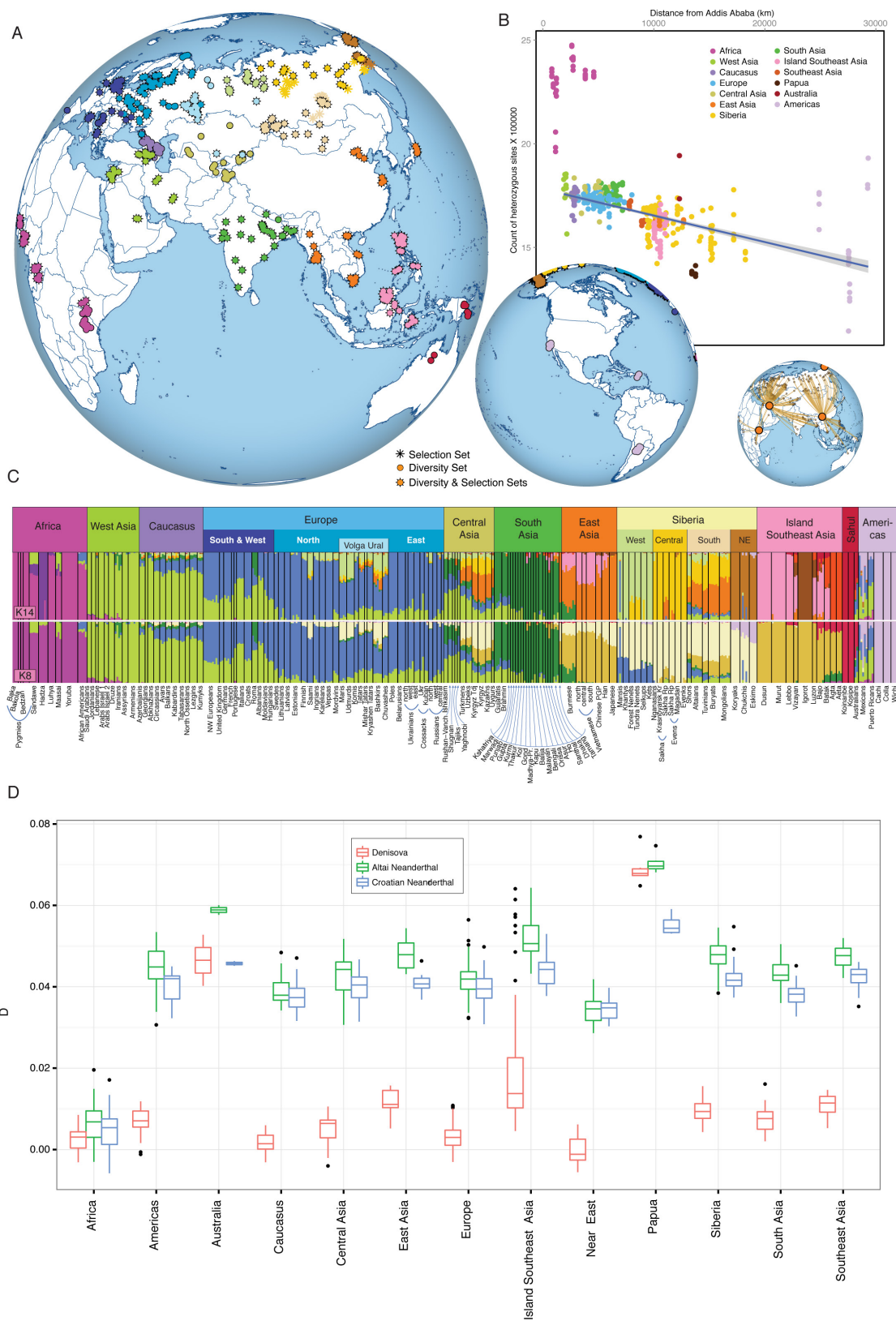
Estimates less than 0 are set to 0. The genome-wide estimate is obtained by weighting each  $\theta$  by the proportion of the genome that was painted with that donor. Neanderthal and Denisova haplotypes were assumed to be proxied by PNG.DEN (0% xOoA by assumption); African haplotypes by PNG.AFR; Papuan and Australian by PNG.PNG and all other haplotypes by PNG.EUR. We obtain confidence intervals by bootstrap resampling of haplotypes for each donor/recipient pair.

We estimate the proportion of xOoA in Papuan haplotypes assigned as both Eurasian (0.1%, 95% CI 0–2.6) and Papuan (4%, 95% CI 2.9–4.5) (Supplementary Information 2.2.10), by using the estimated mutation density in xOoA.

**Y chromosome and mtDNA haplogroup analysis.** The presence of an extinct xOoA trace in the genome of modern Papuans may seem at odds with analyses of mtDNA and Y chromosome phylogenies, which point to a single, recent origin for all non-African lineages (mtDNA L3, which gives rise to all mtDNA lineages outside Africa has been dated at ~70,000 years old<sup>33,34</sup>). However, uniparental markers inform on a small fraction of our genetic history, and a single origin for all non-African lineages does not exclude multiple waves OoA from a shared common ancestor. We show analytically (Supplementary Information 2.2.12) that, if the xOoA signature entered the genome of Papuan individuals > 40 kya, their mtDNA and Y lineages could have been lost by genetic drift even assuming an initial xOoA mixing component of up to 35%. Similar findings have been reported recently<sup>13</sup>.

1. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
2. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).
3. Behar, D. M. *et al.* A “Copernican” reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90**, 675–684 (2012).
4. Soares, P. *et al.* The archaeogenetics of Europe. *Curr. Biol.* **20**, R174–R183 (2010).

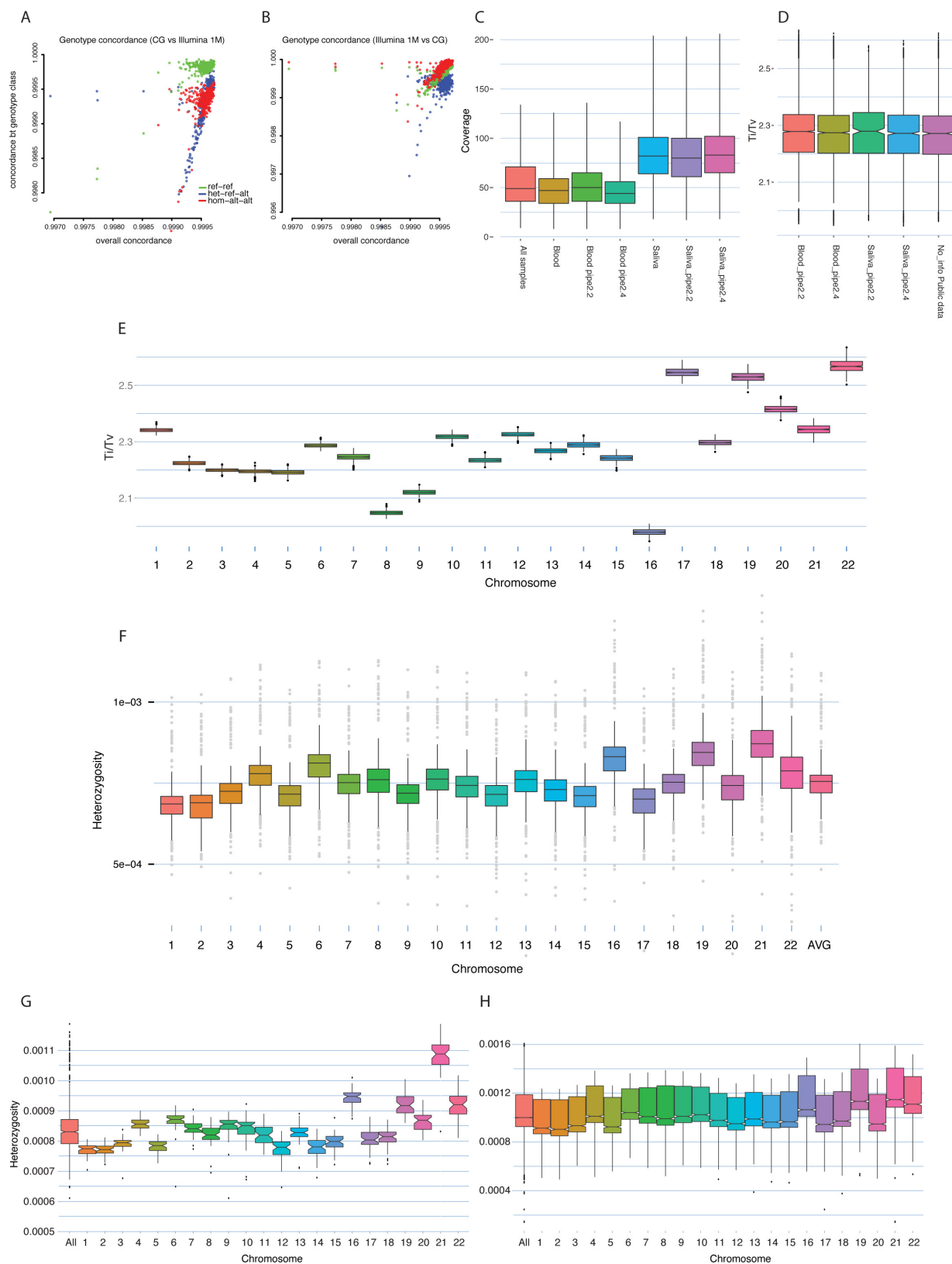




### Extended Data Figure 1 | Sample Diversity and Archaic signals.

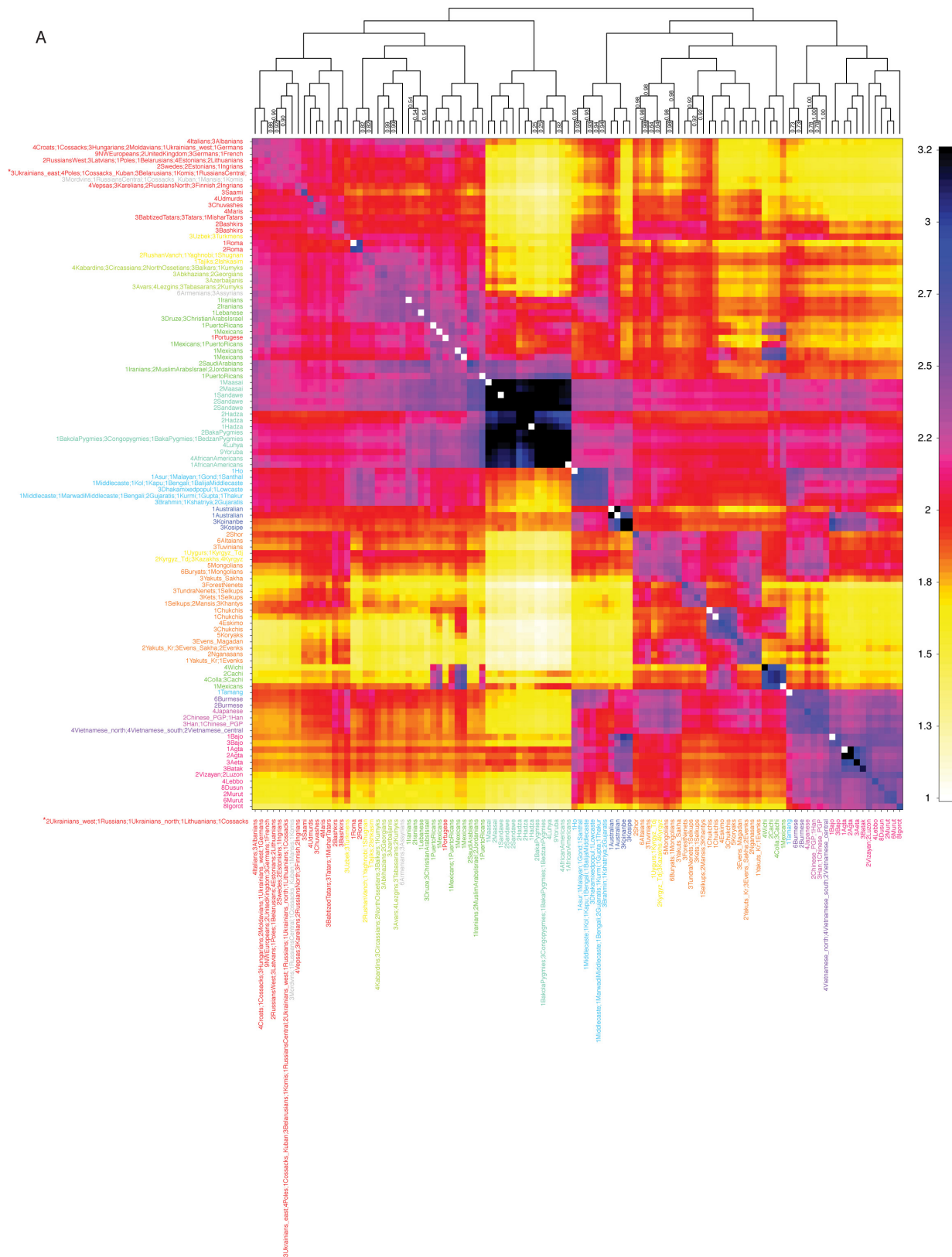
**a**, Map of location of samples highlighting the diversity/selection sets.  
**b**, Sample-level heterozygosity is plotted against distance from Addis Ababa. The trend line represents only non-African samples. The inset shows the waypoints used to arrive at the distance in kilometres for each sample.  
**c**, ADMIXTURE plot ( $K = 8$  and  $14$ ) which relates general visual inspection of genetic structure to studied populations and their region of

origin. **d**, Box plots were used to visualize the Denisova (red), Altai (green) and Croatian Neanderthal (blue)  $D$  distribution for each regional group of samples. Oceanian Altai  $D$  values show a remarkable similarity with the Denisova  $D$  values for the same region, in contrast with the other groups of samples where the Altai box plots tend to be more similar to the Croatian Neanderthal ones. Boxes show median, first and third quartiles, with 1.5 $\times$  interquartile range whiskers and black dots as outliers.



**Extended Data Figure 2 | Data quality checks and heterozygosity patterns.** **a, b**, Concordance of DNA sequencing (Complete Genomics Inc.) and DNA genotyping (Illumina genotyping arrays) data (ref-ref; het-ref-alt and hom-alt-alt, see Supplementary Information 1.6) from chip (**a**) and sequence data (**b**). **c**, Coverage (depth) distribution of variable positions, divided by DNA source (blood or saliva) and complete genomic calling pipeline (release version). **d**, Genome-wide distribution of transition/transversion ratio subdivided by DNA source (saliva or blood) and by complete genomic calling pipeline. **e**, Genome-wide

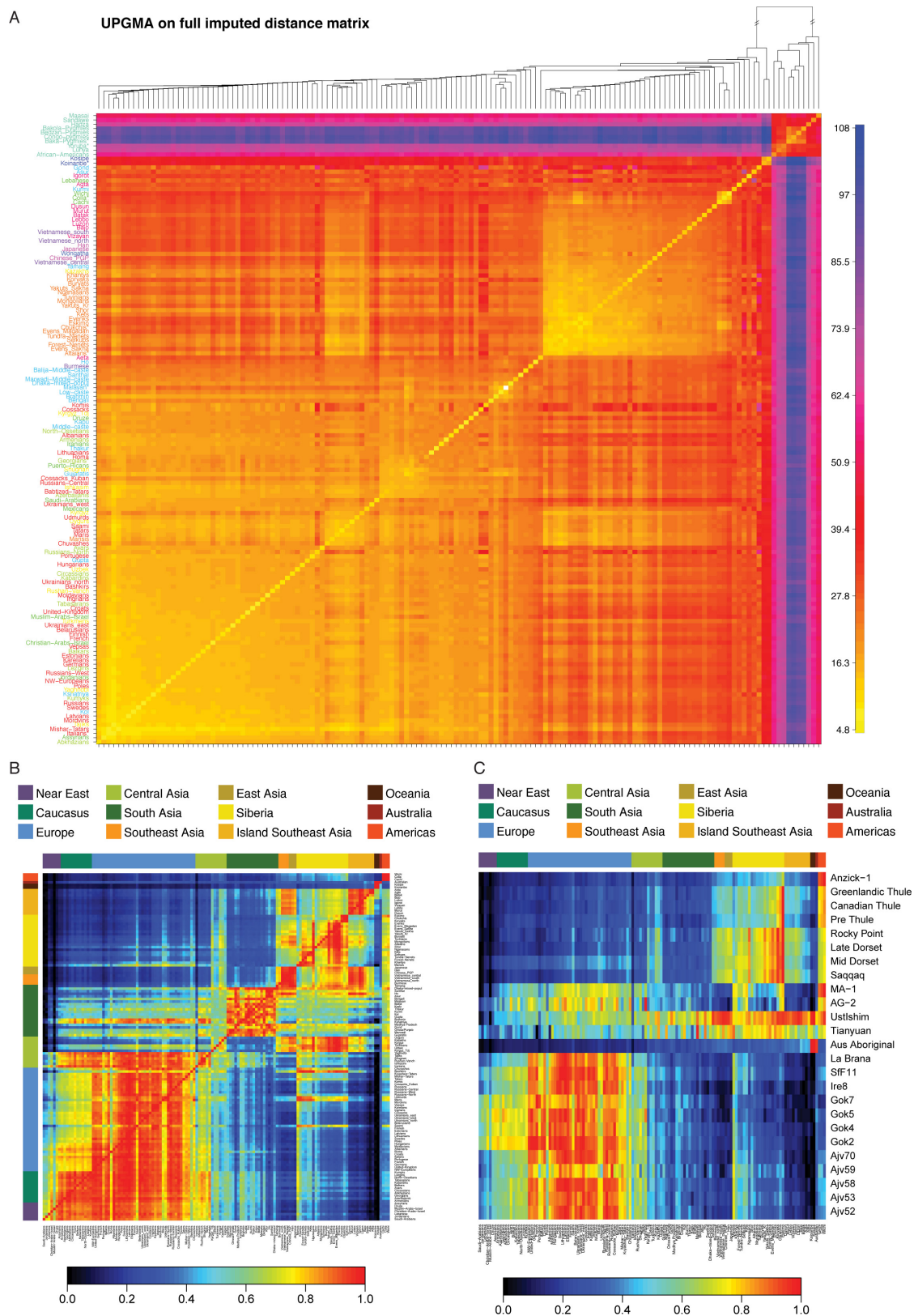
distribution of transition/transversion ratio subdivided by chromosomes. **f**, Inter-chromosome differences in observed heterozygosity in 447 samples from the diversity set. **g**, Inter-chromosome differences in observed heterozygosity in a set of 50 unpublished genomes from the Estonian Genome Center, sequenced on an Illumina platform at an average coverage exceeding 30×. **h**, Inter-chromosome differences in observed heterozygosity in the phase 3 of the 1000 Genomes Project. The total number of observed heterozygous sites was divided by the number of accessible base pairs reported by the 1000 Genomes Project.



**Extended Data Figure 3 | FineSTRUCTURE shared ancestry analysis.** ChromoPainter and FineSTRUCTURE results, showing both inferred populations and the underlying (averaged) number of haplotypes that an individual in a population receives (rows) from donor individuals in other populations (columns). 108 populations are inferred by FineSTRUCTURE. The dendrogram shows the inferred relationship between populations.

The numbers on the dendrogram give the proportion of MCMC iterations for which each population split is observed (where this is less than 1). Each 'geographical region' has a unique colour from which individuals are labelled. The number of individuals in each population is given in the label; for example, '4Italians; 3Albanians' is a population of size 7 containing 4 individuals from Italy and 3 from Albania.

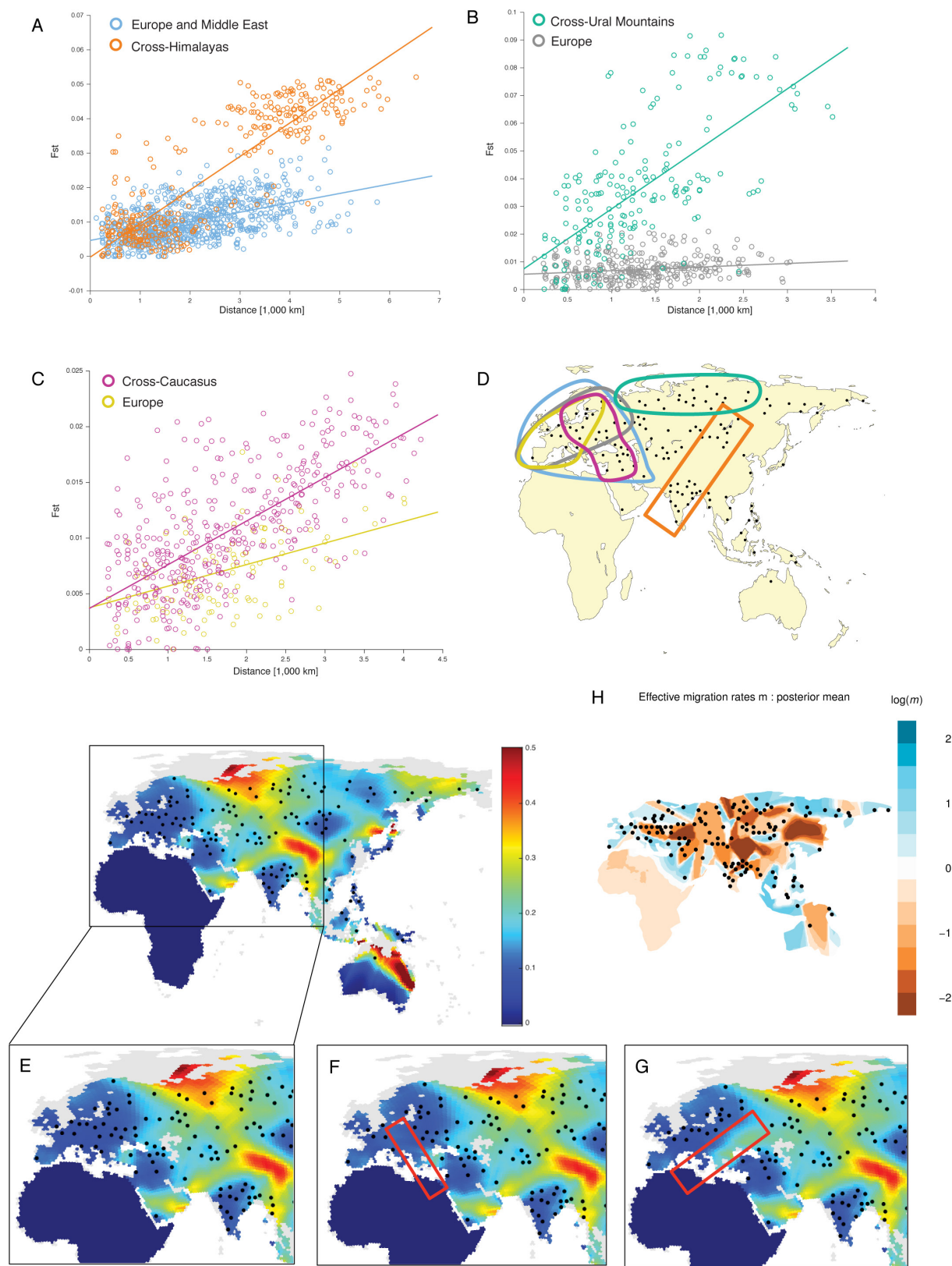




Extended Data Figure 4 | See next page for caption.

**Extended Data Figure 4 | MSMC genetic split times and outgroup  $f_3$  results.** **a**, The MSMC split times estimated between each sample and a reference panel of nine genomes were linearly interpolated to infer the broader square matrix. **b, c**, Summary of outgroup  $f_3$  statistics for each pair of non-African populations or an ancient sample using Yoruba as an outgroup. Populations are grouped by geographic region and are ordered with increasing distance from Africa (left to right for columns and bottom to top for rows). Colour bars at the left and top of the heat map indicate the colour coding used for the geographical region. Individual population labels are indicated at the right and bottom of the heat map. The  $f_3$  statistics are scaled to lie between 0 and 1, with a black colour indicating those close to 0 and a red colour indicating those close to 1. Let  $m$  and  $M$  be the minimum and maximum  $f_3$  values within a

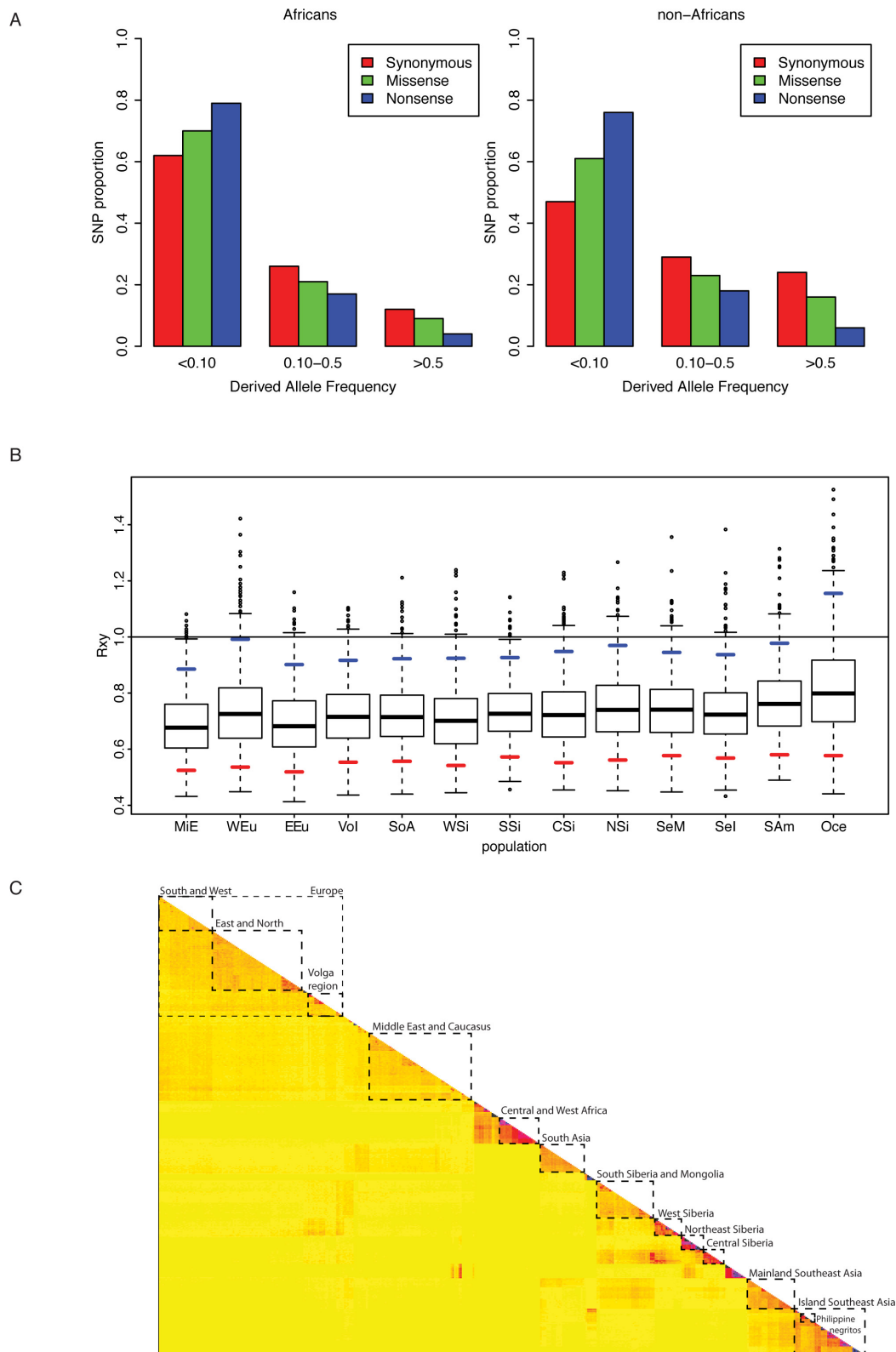
given row (that is, focal population). That is, for focal population  $X$  (on rows),  $m = \min_{Y, Y \neq X} f_3(X, Y; \text{Yoruba})$  and  $M = \max_{Y, Y \neq X} f_3(X, Y; \text{Yoruba})$ . The scaled  $f_3$  statistic for a given cell in that row is given by  $f_{3\text{scaled}} = (f_3 - m)/(M - m)$ , so that the smallest  $f_3$  in the row has value  $f_{3\text{scaled}} = 0$  (black) and the largest has value  $f_{3\text{scaled}} = 1$  (red). By default, the diagonal has value  $f_{3\text{scaled}} = 1$  (red). The heat map is therefore asymmetric, with the population closest to the focal population at a given row having value  $f_{3\text{scaled}} = 1$  (red colour) and the population farthest from the focal population at a given row having value  $f_{3\text{scaled}} = 0$  (black colour). Therefore, at a given row, scanning the columns of the heat map reveals the populations with the most shared ancestry with the focal population of that row in the heat map.



**Extended Data Figure 5 | Geographical patterns of genetic diversity.** Isolation by distance pattern across areas of high genetic gradient, using Europe as a baseline. The samples used in each analysis are indicated by coloured lines on the maps to the right of each plot. **a–d**, The panels show  $F_{ST}$  as a function of distance across the Himalayas (**a**), the Ural

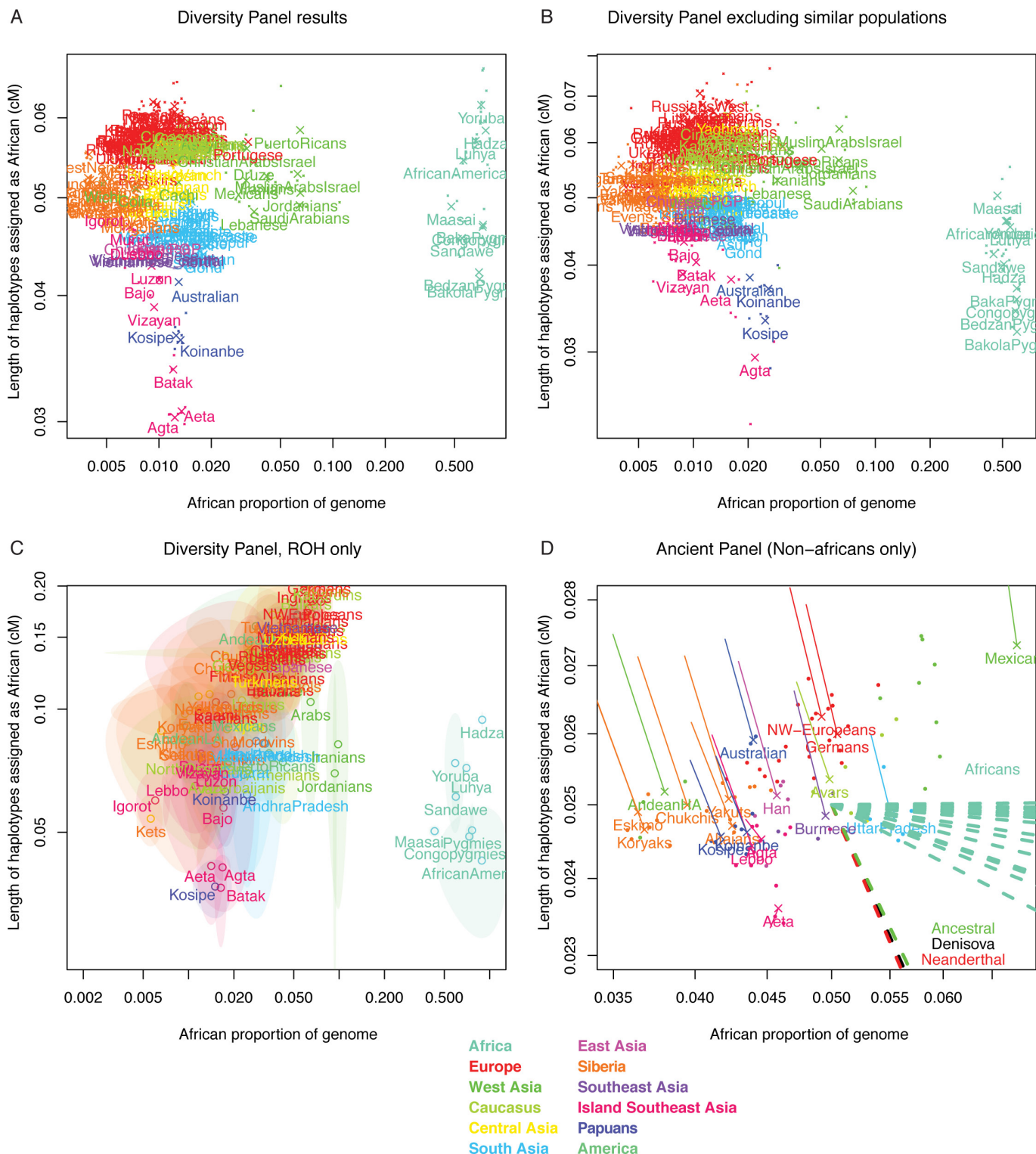
mountains (**b**), and the Caucasus (**c**) as reported on the colour-coded map (**d**). **e**, Effect of creating gaps in the samples in Europe. **f, g**, We tested the effect of removing samples from stripes, either north to south (**f**) or west to east (**g**), to create gaps comparable in size to the gaps in samples in the dataset. **h**, Effective migration surfaces inferred by EEMS.





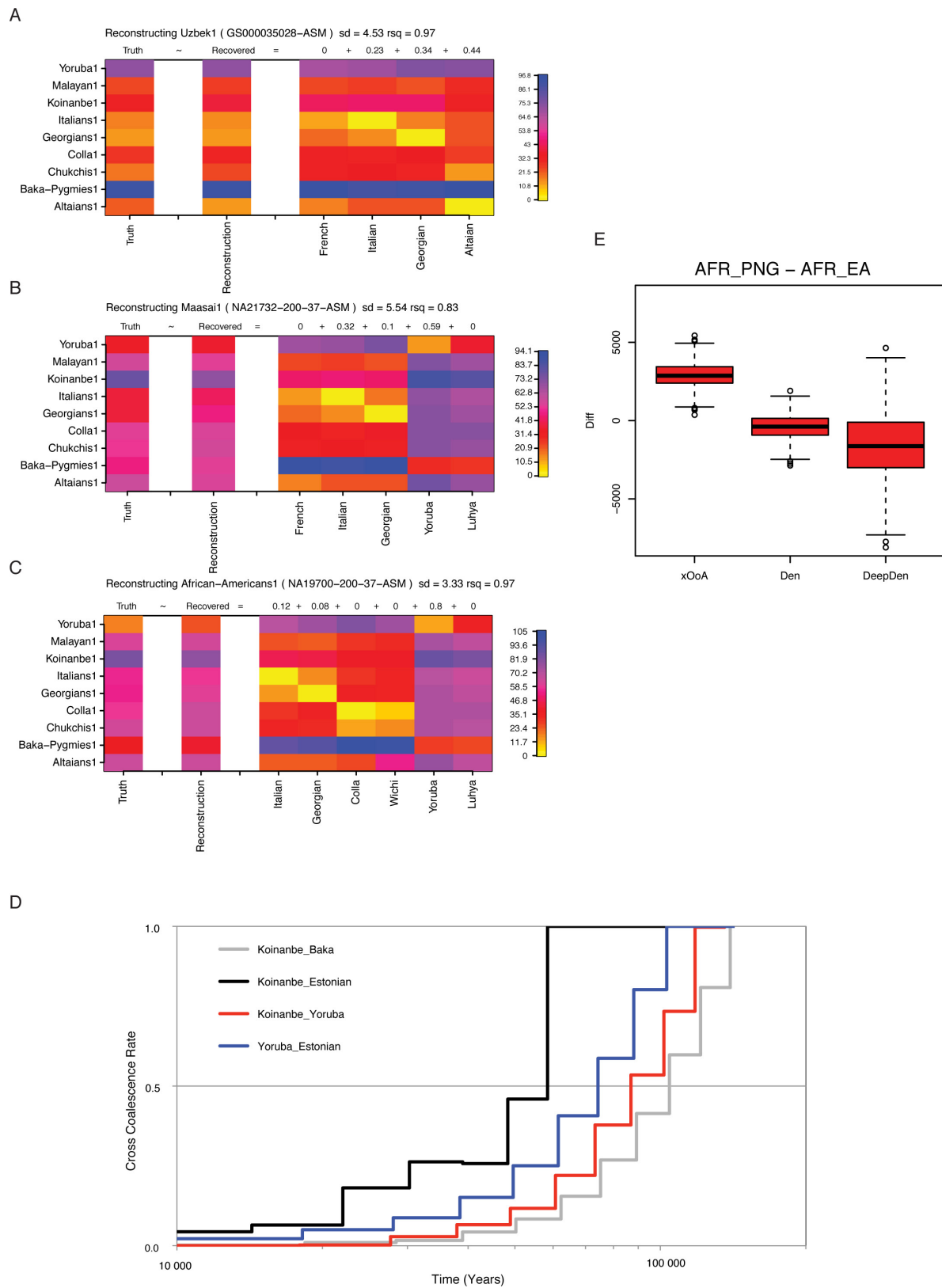
**Extended Data Figure 6 | Summary of positive selection results. a,** Bar plot comparing frequency distributions of functional variants in Africans and non-Africans. The distribution of exonic SNPs according to their functional impact (synonymous, missense and nonsense) as a function of allele frequency. Note that the data from both groups was normalized for a sample size of  $n = 21$  and that the Africans show significantly ( $\chi^2 P < 1 \times 10^{-15}$ ) more rare variants across all sites classes. **b,** Result of 1,000 bootstrap replica of the  $R_{X/Y}$  test for a subset of pigmentation genes highlighted by Genome Wide Association Studies (GWAS,  $n = 32$ ). The horizontal line provides the African reference ( $x = 1$ ) against which all

other groups are compared. The blue and red marks show the 95th and the 5th percentile of the bootstrap distributions respectively. If the 95th percentile is below 1, then the population shows a significant excess of missense variants in the pigmentation subset relative to the Africans. Note that this is the case for all non-Africans except the Oceanians. **c,** Pools of individuals for selection scans. fineSTRUCTURE-based co-ancestry matrix was used to define twelve groups of populations for the downstream selection scans. These groups are highlighted in the plot by boxes with broken line edges. The number of individuals in each group is reported in Supplementary Table 1:3.2-1.



**Extended Data Figure 7 | Length of haplotypes assigned as African by fineSTRUCTURE as a function of genome proportion.** a, 447 Diversity Panel results, showing label averages (large crosses) along with individuals (small dots). b, Relative excluded Diversity Panel results, to check for whether including related individuals affects African genome fraction. Individuals that shared more than 2% of genome fraction were forbidden from receiving haplotypes from each other, and the painting was re-run on a large subset of the genome (all run of homozygosity (ROH) regions from any individual). c, ROH-only African haplotypes. To guard against phasing errors, we analysed only regions for which an individual was in a long (>500 kb) run of homozygosity using the PLINK command ‘-homozyg-window-kb 500000-homozyg-window-het 0-homozyg-

density 10’. Because there are so few such regions, we report only the population average for populations with two or more individuals, as well as the standard error in that estimate. Populations for which the 95% confidence interval passed 0 were also excluded. Note the logarithmic axis. d, Ancient DNA panel results. We used a different panel of 109 individuals which included three ancient genomes. We painted chromosomes 11, 21 and 22 and report as crosses the population averages for populations with two or more individuals. The solid thin lines represent the position of each population when modern samples only are analysed. The dashed lines lead off the figure to the position of the ancient hominins and the African samples.



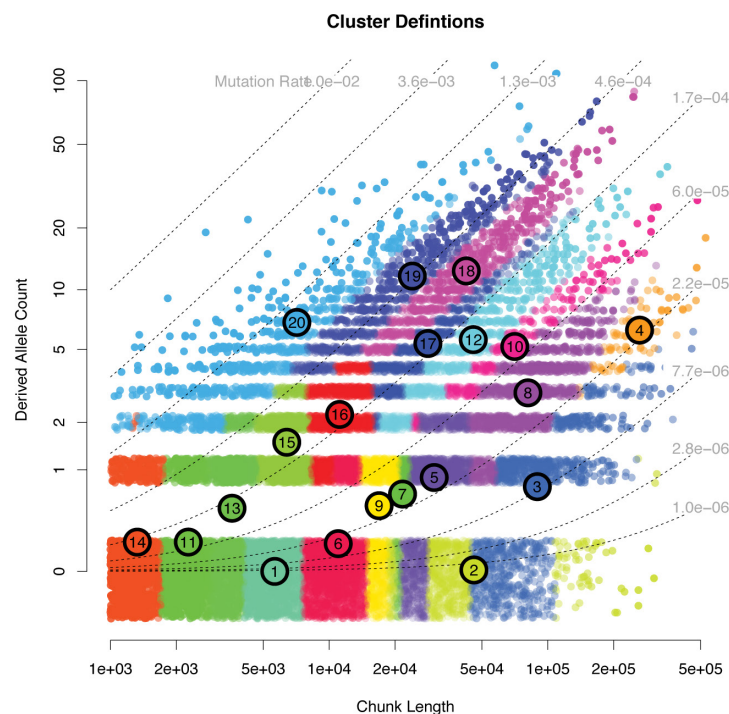
Extended Data Figure 8 | See next page for caption.



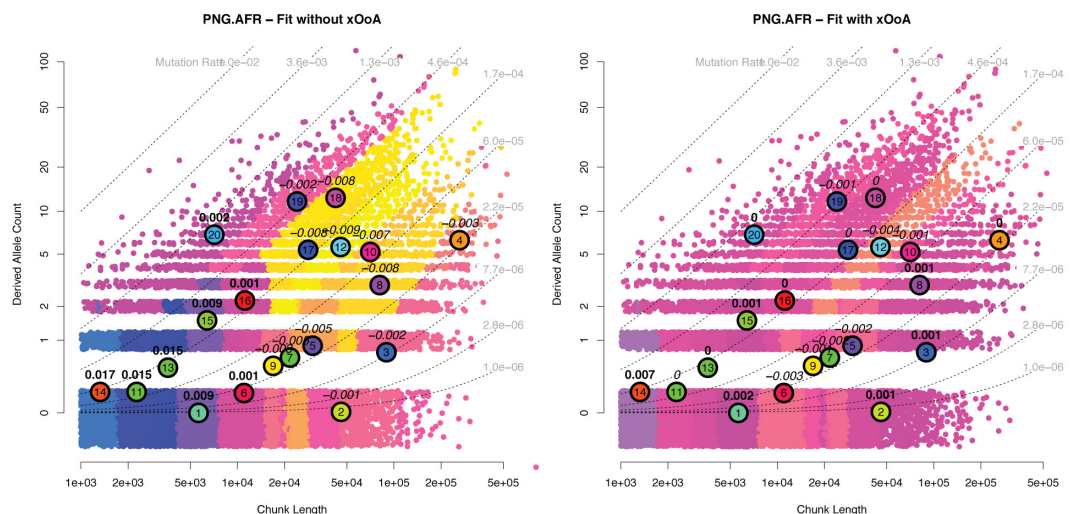
**Extended Data Figure 8 | MSMC Linear behaviour of MSMC split estimates in presence of admixture.** **a–c**, The examined Central Asian (**a**), East African (**b**), and African–American (**c**) genomes yielded a signature of MSMC split time (truth, left-most column) that could be recapitulated (reconstruction, second left-most column) as a linear mixture of other MSMC split times. The admixture proportions inferred by our method (top of each admixture component column) were remarkably similar to the ones previously reported from the literature. **d**, MSMC split times calculated after re-phasing an Estonian and a Papuan (Koinanbe) genome together with all the available West African and Pygmy genomes from our dataset to minimize putative phasing artefacts. The cross coalescence rate curves reported here are quantitatively comparable with the ones of Fig. 2a, hence showing that phasing artefacts are unlikely to explain

the observed past-ward shift of the Papuan–African split time. **e**, Box plot showing the distribution of differences between African–Papuan and African–Eurasian split times obtained from coalescent simulations assembled through random replacement to make 2,000 sets of 6 individuals (to match the 6 Papuans available from our empirical dataset), each made of 1.5 Gb of sequence. The simulation command line used to generate each chromosome made of 5 Mb was as follows, where  $x$  is the variable for the divergence time used.  $x = 0.064, 0.4$  or  $0.8$  for the xOoA, Denisova (Den) and Divergent Denisova (DeepDen) cases, respectively. `ms0ancient2 10 1. 065.05 -t 5000. -r 3000. 5000000 -I 7 1 1 1 1 2 2 2 -en 0. 1 .2 -en 0. 2 .2 -en 0. 3 .2 -en 0. 4 .2 -es .025 7.96 -en .025 8.2 -ej.03 7 6 -ej.04 6 5 -ej.060 8 3 -ej.061 4 3 -ej.062 2 1 -ej.063 3 1 -ej x 1 5.`

A



B

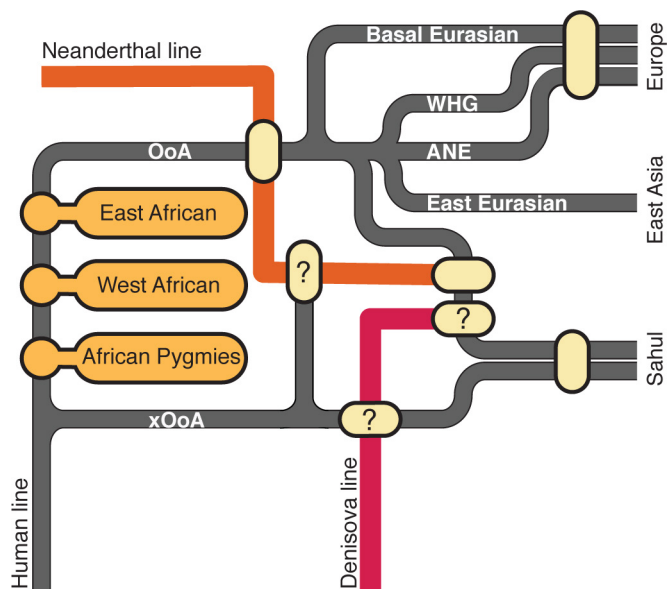


C

	PNG.DEN	EUR.DEN	PNG.AFR	EUR.AFR	xOoA
Average Length	19518	17554	28499	32824	17783
Non-African Derived allele mutation rate	0.000202	0.000142	0.000061	0.000049	0.000074
Relative age to OoA (mutational clock)	4.11	2.88	1.24	1	1.51

**Extended Data Figure 9 | Modelling the xOoA components with FineSTRUCTURE.** **a**, Joint distribution of haplotype lengths and derived allele count, showing the median position of each cluster and all haplotypes assigned to it in the maximum a posteriori (MAP) estimate. Note that although a different proportion of points is assigned to each in the MAP, the total posterior is very close to  $1/K$  for all. The dashed lines show a constant mutation rate. Haplotypes are ordered by mutation rate from low to high. **b**, Residual distribution comparison between the two-component mixture using EUR.AFR and EUR.PNG (left), and the

three-component mixture including xOoA (using the same colour scale) (right). The root mean square error (RMSE) residuals without xOoA are larger ( $RMSE = 0.0055$  compared to  $RMSE = 0.0018$ ) but more importantly, they are also structured. **c**, Assuming a mutational clock and a correct assignment of haplotypes, we can estimate the relative age of the splits from the number of derived alleles observed on the haplotypes. This leads to an estimate of 1.5 times older for xOoA compared to the Eurasian–Africa split.



**Extended Data Figure 10 | Proposed xOoA model.** A schematic illustrating, as suggested by the results presented here, a model of an early, extinct Out-of-Africa (xOoA) signature in the genomes of Sahul populations at their arrival in the region. Given the overall small genomic contribution of this event to the genomes of modern Sahul individuals, we could not determine whether the documented Denisova admixture (question marks) and putative multiple Neanderthal admixtures took place along this extinct OoA. We also speculate (question mark) people who migrated along the xOoA route may have left a trace in the genomes of the Altai Neanderthal as reported by Kuhlwiilm and colleagues<sup>12</sup>.

# *De novo* assembly and phasing of a Korean human genome

Jeong-Sun Seo<sup>1,2,3,4,5\*</sup>, Arang Rhie<sup>1,2,3\*</sup>, Junsoo Kim<sup>1,4\*</sup>, Sangjin Lee<sup>1,5\*</sup>, Min-Hwan Sohn<sup>1,2,3</sup>, Chang-Uk Kim<sup>1,2,3</sup>, Alex Hastie<sup>6</sup>, Han Cao<sup>6</sup>, Ji-Young Yun<sup>1,5</sup>, Jihye Kim<sup>1,5</sup>, Junho Kuk<sup>1,5</sup>, Gun Hwa Park<sup>1,5</sup>, Juhyeok Kim<sup>1,5</sup>, Hanna Ryu<sup>4</sup>, Jongbum Kim<sup>4</sup>, Mira Roh<sup>4</sup>, Jeonghun Baek<sup>4</sup>, Michael W. Hunkapiller<sup>7</sup>, Jonas Korch<sup>7</sup>, Jong-Yeon Shin<sup>1,5</sup> & Changhoon Kim<sup>4</sup>

Advances in genome assembly and phasing provide an opportunity to investigate the diploid architecture of the human genome and reveal the full range of structural variation across population groups. Here we report the *de novo* assembly and haplotype phasing of the Korean individual AK1 (ref. 1) using single-molecule real-time sequencing<sup>2</sup>, next-generation mapping<sup>3</sup>, microfluidics-based linked reads<sup>4</sup>, and bacterial artificial chromosome (BAC) sequencing approaches. Single-molecule sequencing coupled with next-generation mapping generated a highly contiguous assembly, with a contig N50 size of 17.9 Mb and a scaffold N50 size of 44.8 Mb, resolving 8 chromosomal arms into single scaffolds. The *de novo* assembly, along with local assemblies and spanning long reads, closes 105 and extends into 72 out of 190 euchromatic gaps in the reference genome, adding 1.03 Mb of previously intractable sequence. High concordance between the assembly and paired-end sequences from 62,758 BAC clones provides strong support for the robustness of the assembly. We identify 18,210 structural variants by direct comparison of the assembly with the human reference, identifying thousands of breakpoints that, to our knowledge, have not been reported before. Many of the insertions are reflected in the transcriptome and are shared across the Asian population. We performed haplotype phasing of the assembly with short reads, long reads and linked reads from whole-genome sequencing and with short reads from 31,719 BAC clones, thereby achieving phased blocks with an N50 size of 11.6 Mb. Haplotigs assembled from single-molecule real-time reads assigned to haplotypes on phased blocks covered 89% of genes. The haplotigs accurately characterized the hypervariable major histocompatibility complex region as well as demonstrating allele configuration in clinically relevant genes such as *CYP2D6*. This work presents the most contiguous diploid human genome assembly so far, with extensive investigation of unreported and Asian-specific structural variants, and high-quality haplotyping of clinically relevant alleles for precision medicine.

Although massively parallel sequencing approaches have been widely used to study genomic variation, simple alignment of short reads to a reference genome cannot be used to investigate the full range of structural variation and phased diploid architecture, which are important for precision medicine. By contrast, the single-molecule real-time (SMRT) sequencing platform produces long reads that can resolve repetitive structures effectively. We integrated this technology with several other sequencing approaches to construct a high-quality Korean diploid genome assembly (Extended Data Fig. 1).

SMRT sequencing of the genome of a Korean individual AK1, for whom we have previously reported the annotated variations assessed with BAC clones and array comparative genomic hybridization<sup>1</sup>, was performed at 101× coverage using Pacific Biosciences (PacBio) RSII

(Extended Data Fig. 2a). Reads were assembled and error-corrected with FALCON and Quiver<sup>5</sup> to generate 3,128 contigs with a contig N50 length of 17.9 Mb (Extended Data Table 1, Extended Data Fig. 2b and Supplementary Tables 1–3). To anchor these contigs into larger scaffolds, we used next-generation mapping (NGM) from BioNano Genomics Irys System, which produces physical maps with unique sequence motifs that can provide long-range structural information of the genome. Two rounds of NGM at 97× and 108× coverage were performed, with the second designed to protect fragments better from breakage at fragile sites, providing improved long-range anchoring (Supplementary Table 4). The optical maps were assembled *de novo* into genome maps. Hybrid scaffolding of the contigs and genome maps resulted in 2,832 scaffolds with a scaffold N50 size of 44.8 Mb (Extended Data Table 1 and Extended Data Fig. 3a). Because NGMs provide orders of magnitude longer range information (Supplementary Table 4) compared to long reads from the SMRT platform (Supplementary Table 1), we relied on the genome map when there were conflicts between the two datasets. Checks for consistency between genome maps and contigs corrected potential assembly errors within 23 contigs (Extended Data Fig. 3b and Supplementary Table 5). The final assembly after polishing with Illumina reads (Extended Data Fig. 4a) is characterized by marked contiguity that has not been achieved by non-reference assemblies of the human diploid genome<sup>6–8</sup> so far, and improves on the previous best<sup>6</sup> N50 length by 18 Mb (Table 1). The largest 91 scaffolds, for example, cover 90% of the genome and 8 chromosomal arms are spanned by single scaffolds (Fig. 1a).

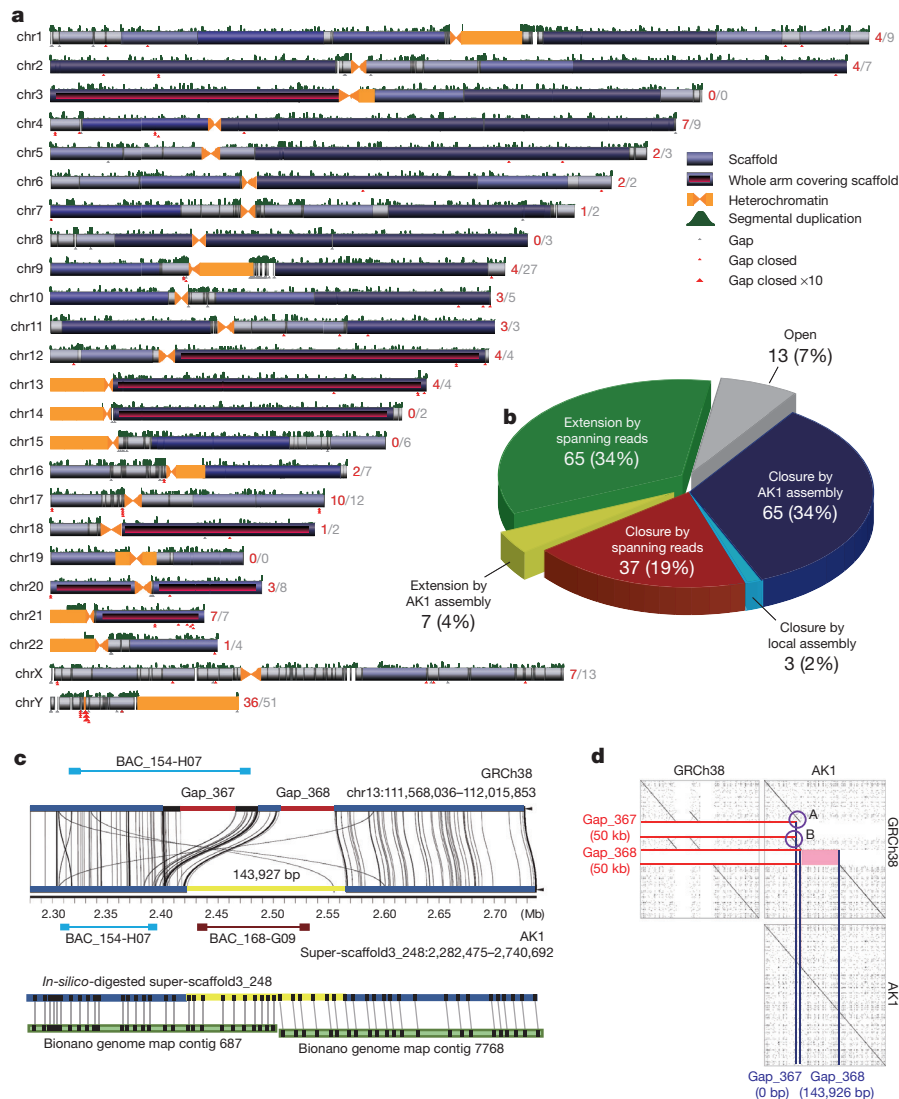
The scaffolding accuracy of the AK1 assembly was assessed using paired-end sequences from AK1 BAC library<sup>1</sup> from 62,758 BAC clones (Extended Data Fig. 1). Most (95.4%) of the uniquely aligned BAC clones were in concordance with the assembly (Extended Data Table 2), as expected since the genomic DNA originated from the same individual. From the set of BAC clones that aligned concordant with the reference genome, 99.8% also aligned concordant with the AK1 assembly, with most of the discrepancies caused by phase differences (Supplementary Table 6). The base accuracy of the assembly was assessed by Illumina short reads (72×). The read-depth distributions of the reads mapped to GRCh37, GRCh38 and AK1 show similar patterns (Extended Data Fig. 4b). The estimated base-level error rate of the assembly was less than 10<sup>−5</sup> based on the count of single nucleotide polymorphisms (SNPs) with unexpected alleles (Extended Data Fig. 4c and Supplementary Table 7).

We used the AK1 assembly to close gaps remaining in human genome reference GRCh38. Of 190 euchromatic gaps (Supplementary Table 8), 65 were closed entirely by our *de novo* assembly (Fig. 1b and Extended Data Fig. 5). Local realignment and reassembly, and use of spanning reads, resolved a further 40 gaps. The closed gaps were

<sup>1</sup>Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul 110-799, South Korea. <sup>2</sup>Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul 110-799, South Korea. <sup>3</sup>Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 110-799, South Korea. <sup>4</sup>Bioinformatics Institute, Macrogen Inc., Seoul 153-023, South Korea. <sup>5</sup>Genome Institute, Macrogen Inc., Seoul 153-023, South Korea. <sup>6</sup>BioNano Genomics, San Diego, California 92121, USA. <sup>7</sup>Pacific Biosciences of California, Inc., Menlo Park, California 94025, USA.

\*These authors contributed equally to this work.





**Figure 1 | AK1 *de novo* assembly scaffolds compared to GRCh38.**

**a**, Scaffold coverage over GRCh38 per chromosome. The blue shading represents scaffold size, with darker segments for longer scaffolds. Eight chromosomal arms are spanned by single scaffolds. Closed euchromatic gaps are labelled in red on each chromosome, with the total number of gaps in grey. **b**, Number of gaps closed using the AK1 assembly (blue), local assembly of long reads (light blue), and long reads alone (red). The number of extended gaps with AK1 assembly is represented in yellow, with long reads in green and open gaps in grey. The 65 dot plots of gaps

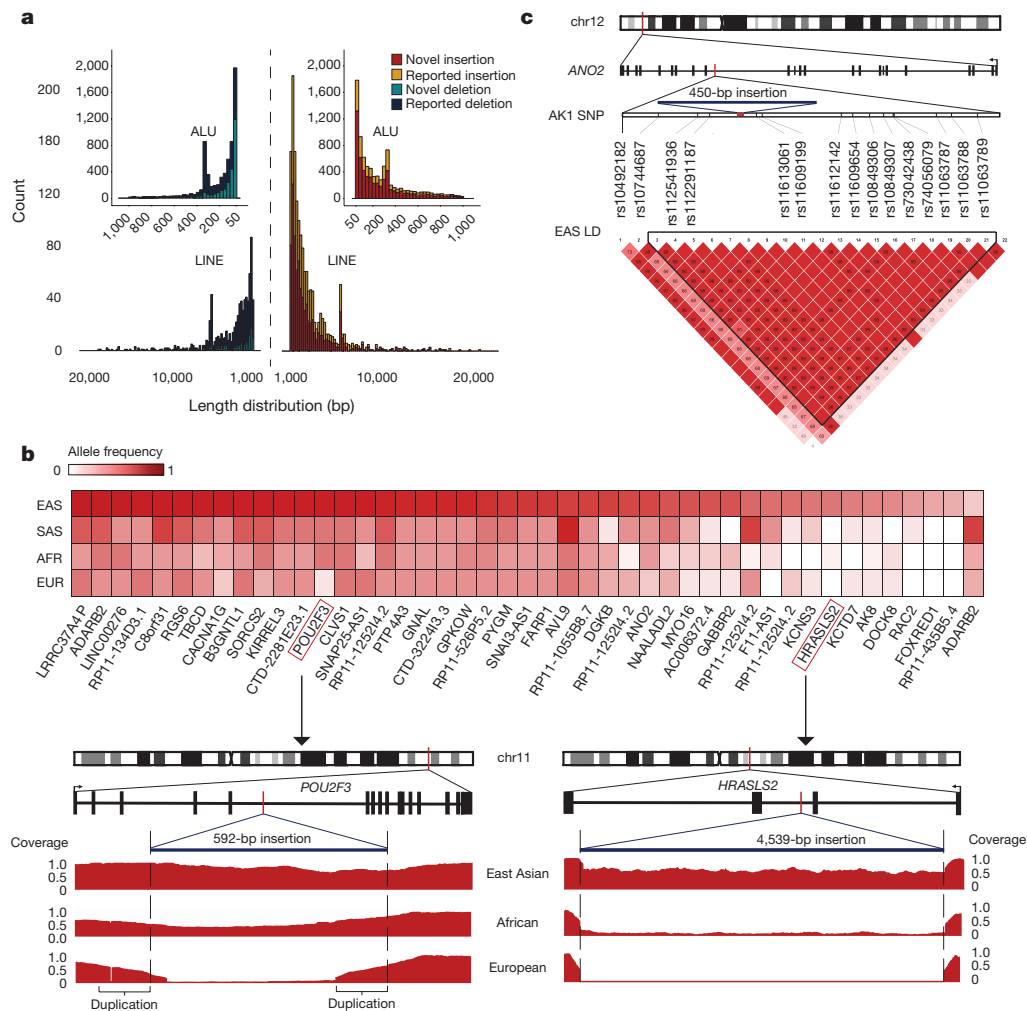
filled with a total of 364 kb of sequence into 1.5 Mb (Supplementary Table 9 and Supplementary Information). We also extended into 72 of the 85 remaining gaps with the addition of 663 kb of sequence into 4.1 Mb. These locations, previously intractable using only short reads, commonly contained simple tandem repeats, as reported previously<sup>6,9</sup>. One example (Fig. 1c, d) illustrates two gaps resolved by AK1 assembly with supporting evidence from BACs and genome maps.

We identified 18,210 structural variants (SVs), including 7,358 deletions, 10,077 insertions, 71 inversions, and 704 complex variants at a base resolution through the direct comparison between the AK1 assembly and the human reference genome GRCh37 (Supplementary Tables 10, 11). We were able to validate 271 out of 276 SVs with BAC contigs generated by SMRT sequencing (Supplementary Table 12). Compared to previous studies<sup>6,8–11</sup>, a total of 11,927 variants were previously unreported, which account for approximately 47% (3,465) and 76% (7,710) of all deletions and insertions, respectively (Fig. 2a and Extended Data Fig. 6a). Of the new SVs, 86% were highly enriched for

closed with the AK1 assembly can be found in the AK1 genome browser (<http://211.110.34.36/gbrowse2>). **c**, AK1 assembly resolving two gaps along with BACs and optical map suggests that gap\_367 and both its edges (red and black bars) shrink to zero, whereas gap\_368 expands to 144 kb (yellow bar). **d**, Three dot plots show how unique sequences have been added to the reference genome. Reference–reference (top left), reference–AK1 assembly (top right) and AK1–AK1 (bottom right). A and B indicate deleted GRCh38 sequence around gap\_367.

clusters of mobile and tandem repeats (Extended Data Fig. 6b). PacBio long-read sequencing of the corresponding transcriptome revealed that 155 isoforms are expressed from 54 novel insertion loci, indicating the existence of functional elements in human genomes that were probably undetectable using short reads (Supplementary Table 13). A total of 4,326 deletions and 5,833 insertions occurred within 6,073 genes. Out of 615 exonic variants, 427 were new, and 68% of them did not affect protein functionality by maintaining the reading frame or occurring within non-protein coding genes. Among the new amino-acid-changing variants, 77% were composed of mobile or tandem repeats (Supplementary Table 14), and functional annotation clustering with the 31 genes, which contain the remaining non-repetitive variants, using DAVID<sup>12</sup> showed that they were predominantly related to ion binding, epidermal growth factor, and fibronectin.

Investigation of the insertions suggested that the AK1 sequences consist not only of repeats and duplications, but also of unique sequences that are not found in the reference genome. To examine



**Figure 2 | AK1 SV distribution and Asian-specific variants.**

**a**, Distribution of insertions (red/orange) and deletions (cyan/dark blue) between AK1 and GRCh37, compared to SVs identified from previous studies. In total, 47% and 76% of the insertions and deletions, respectively, were previously unreported. **b**, Allele frequency of 45 Asian-specific insertions ( $\geq 0.3$  allele frequency difference;  $\leq 0.5$  non-Asian allele frequency). The coverage for the genic insertions was calculated from

38 whole-genome high-coverage samples by dividing the read depth by the median genome coverage across individuals with the same ancestry. **c**, In *ANO2*, the Asian-specific insertion occurs within an East Asian (EAS) linkage disequilibrium (LD) block, sharing a similar population allele frequency with the adjacent AK1 SNPs. AFR, African; EUR, European; SAS, South Asian.

whether the unique sequences are universal or ancestry specific, we aligned raw reads from high-coverage 1000 Genomes Project samples<sup>10,11</sup> and additional high-coverage Asian samples against our AK1 assembly, and compared the normalized read depths between four ancestral groups. Out of 853 insertions, encompassing 1.7 Mb, which were found in all of the ancestral groups, 800 insertions were also called from the variant analysis with respect to GRCh38, and as such are candidates for addition to the human reference genome (Supplementary Tables 15, 16). Moreover, 400 insertions showed highly polymorphic frequency variability across the populations, and 76 of them, including 45 genic insertions, were Asian specific. Among the genic insertions, we found that a 592-bp insertion within *POU2F3*, reported to have distinctly variable haplotype frequencies among populations<sup>13</sup>, was comprised of 452 bp of unique sequence between two 140-bp duplications (Fig. 2b). We also identified numerous large insertions with higher frequency in the Asian population, such as a 4,539-bp insertion in *HRASLS2*. Next, we investigated the haplotype structures associated with Asian-specific variants by using linkage disequilibrium blocks inferred from 1000 Genomes Project Asian samples<sup>10</sup>. Among the variants, 39 insertions were present within the blocks, and 82% of them were located on the same block as a homozygous AK1 SNP, the frequency of which was highest in the Asian

population (Supplementary Table 17). One of the insertions, found within *ANO2*, had a similar allele frequency with adjacent homozygous AK1 SNPs within the same linkage disequilibrium block, suggesting that the insertion shares a single ancient haplotype with the SNPs (Fig. 2c). Our findings demonstrate the important genomic differences of Asian ancestral group from the others, and highlight the need for further genomic studies focused on individuals outside of European ancestry to describe the full range of functionally important variations in humans.

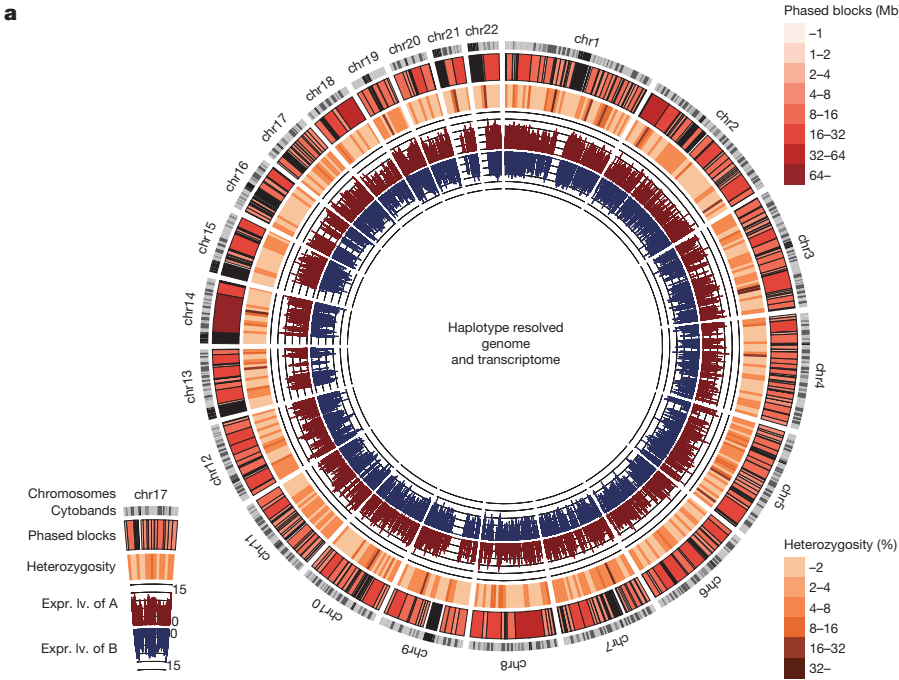
To reflect the diploid genome structure better, we built separate *de novo* assemblies (haplotigs) representing the two haplotypes of each homologous chromosome pair<sup>14</sup>. Phasing was performed with PacBio long reads, Illumina short reads, 10X Genomics linked reads<sup>4</sup> (30 $\times$ ), and reads from BACs representing a single haplotype (47 $\times$ ). Heterozygous SNVs called from these methods are unambiguously assigned to two alternative phases, producing phased blocks with an N50 length of 11.6 Mb, considerably longer than previously reported<sup>4,6,8,15,16</sup> (Table 1). We assessed the accuracy of the phased blocks against the end sequences of BACs, and found a long-range switch error rate to be under 0.3%. SMRT reads were then partitioned into the two phases in which sufficient marker SNVs were present. The two partitioned read sets were assembled *de novo* into haplotigs (Table 1 and Extended Data Table 3).

Table 1 | Comparison of human *de novo* assembly and haplotype phasing summary statistics

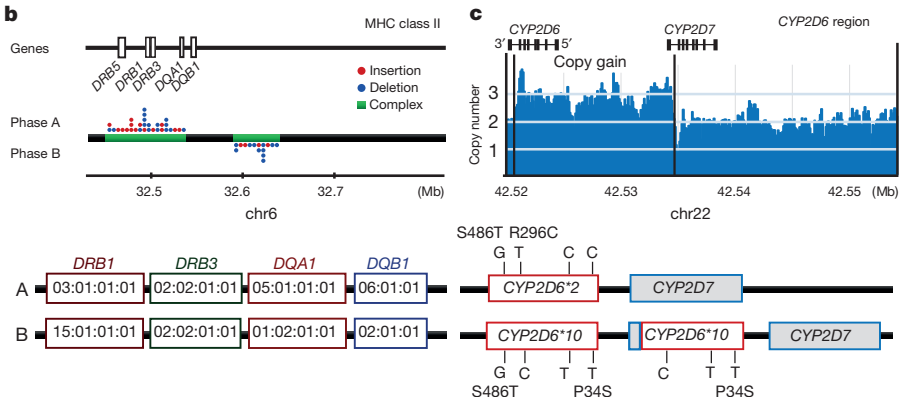
	AK1	HuRef	YH_2.0	NA12878	GRCh38
Assembly approach	WGS and BAC	WGS	WGS and fosmid	WGS	BAC and fosmid
Sequencing and physical mapping	PacBio and BioNano	Sanger	Illumina and CG	PacBio and BioNano	Sanger, FISH, OM and fingerprint contigs
<i>De novo</i> assembly algorithm	FALCON	Celera	SOAPdenovo2	Celera and FALCON	Multiple methods
Phasing approach	<i>De novo</i>	Reference-guided	<i>De novo</i>	Reference-guided	NA
Scaffold/contig N50 (Mb)	44.85/17.92	17.66/0.11	20.52/0.02	26.83/1.56	67.79/56.41
Scaffold/contig L50	21/50	48/7,164	39/40,005	37/532	16/19
No. of scaffolds/contigs	2,832/4,206	4,530/71,333	125,643/361,157	18,903/21,235	735/1,385
No. of gaps	264*	68,109†	235,514†	2,332*	999†
Total gap length (Mb)	37.34	34.43	105.20	146.35	159.97
Total bases/non-N bases in assembly (bp)	2,904,207,288 /2,866,687,809	2,844,000,504 /2,809,571,127	2,911,235,363 /2,806,031,133	3,176,574,379 /3,030,222,093	3,209,286,105 /3,049,316,098
Phased block N50 (Mb)	11.55	0.35	NA	0.15	NA
No. of haplotigs	18,964	NA	24,597	NA	NA
Haplotig N50 (kb)	875	NA	484	NA	NA
Haplotig sum (bp)	4,804,460,182	NA	5,152,727,603	NA	NA

We compared the sequencing platform, algorithms, assembly and phasing statistics of human assemblies so far. The comparison demonstrates the power of single-molecule technologies to generate assemblies with superior assembly statistics than that achieved by short-read sequencing. The assembly statistics were obtained from the NCBI and if the summary statistics were not available from NCBI, the numbers were directly acquired from relevant papers. The accession numbers for HuRef<sup>7</sup>, YH\_2.0 (ref. 8), NA12878 (ref. 6) and GRCh38 assemblies are GCA\_000002125.2, GCA\_000004845.2, GCA\_001013985.1 and GCA\_000001405.15, respectively. CG, complete genomics; FISH, fluorescent *in-situ* hybridization; NA, not applicable; OM, optical mapping; WGS, whole-genome shotgun.

\*Number of spanned gaps.  
†Number of spanned and unspanned gaps.



**Figure 3 | Circular visualization of phased blocks with phase-specific expression and two phased regions of MHC class II and *CYP2D6*.** **a**, Genome-wide map of highly heterozygous regions and expression levels of haplotype A and B in log scale. **b**, HLA genes in the MHC class II region. This highly variable, complex region contained many SVs, making it difficult to phase against the reference genome, but allowed full resolution through the *de novo* approach. For detailed comparison, see Extended Data Fig. 7. **c**, Both haplotypes of *CYP2D6* and *CYP2D7*. A duplicated copy of *CYP2D6* was fused with the last exon of *CYP2D7* on haplotype B.



Comparison of the haplotigs to the human reference led to identification of haplotype-specific alleles including SNPs, short indels and SVs (Supplementary Table 18). In addition to the SVs called from the assembly, 13,436 heterozygous haplotype-specific SVs were identified from haplotigs. We tested the accuracy of these SVs against BAC contigs on the same phase, and found that 67 out of the 69 that could be assessed matched perfectly (Supplementary Table 19). The combined length of SNVs, indels and SVs that were heterozygous between the two haplotigs was 69.8 Mb. Moreover, we were able to measure the expression level from each haplotype genome widely (Fig. 3a).

We examined the haplotypes of human leukocyte antigen (HLA) genes in detail, and confirmed the haplotypes using targeted SMRT sequencing (Supplementary Table 20). To avoid common problems<sup>17</sup> associated with hyperpolymorphic patterns of allelic variation, major histocompatibility complex (MHC) class I and II regions were assembled independently. The MHC class II region was phased successfully despite a large number of SVs, highlighting the utility of our *de novo* phasing approach (Fig. 3b and Extended Data Fig. 7). Our approach also allowed a clinically important duplication of *CYP2D6* to be detected and assigned to one phase (Fig. 3c). This result demonstrates that *de novo* assembly-based phasing has advantages in resolving challenging hypervariable regions, and could be used further for pharmacogenomics (Supplementary Discussion).

Allelic configuration is also particularly important for recessive traits. For example, we were able to phase two genes that contained more than two nonsynonymous, heterozygous alleles known to be associated with recessive diseases (Supplementary Table 21). Variants in *MEFV*<sup>18</sup> and *ADAMTS13* (ref. 19), which are predicted to cause familial Mediterranean fever and Upshaw–Shalman syndrome under the autosomal recessive inheritance pattern, respectively, were found in *cis* configuration, with the partner haplotype left intact.

These results demonstrate the power of *de novo* genome assembly and phasing by integrating SMRT sequencing, genome maps, linked reads and BACs for the generation of high-quality contiguous scaffolds, the detection of the full range of SVs, and for understanding the haplotype structure in clinically relevant genes for precision medicine.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 19 April; accepted 15 September 2016.**

**Published online 5 October 2016.**

- Kim, J.-I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
- Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
- Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nat. Biotechnol.* **30**, 771–776 (2012).
- Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
- Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
- Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
- Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
- Cao, H. *et al.* *De novo* assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33**, 617–622 (2015).

- Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
- Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44–57 (2009).
- Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
- Makoff, A. J. & Flomen, R. H. Detailed analysis of 15q11-q14 sequence corrects errors and gaps in the public access sequence to fully reveal large segmental duplications at breakpoints for Prader–Willi, Angelman, and inv dup(15) syndromes. *Genome Biol.* **8**, R114 (2007).
- Suk, E.-K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21**, 1672–1685 (2011).
- Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
- Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the *de novo* assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).
- Bernot, A. *et al.* Non-founder mutations in the *MEFV* gene establish this gene as the cause of familial Mediterranean fever (FMF). *Hum. Mol. Genet.* **7**, 1317–1325 (1998).
- Kokame, K. *et al.* Mutations and common polymorphisms in *ADAMTS13* gene responsible for von Willebrand factor-cleaving protease activity. *Proc. Natl Acad. Sci. USA* **99**, 11902–11907 (2002).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank J.-I. Kim, J. Sung and T. Bleazard for discussion and assistance with manuscript preparation. We thank M. Boitano and J. Chin for assistance with library preparation and test sequencing runs, and assistance with the FALCON assembler, respectively. We would like to send additional thanks to 10X Genomics for their technical supports. This work has been supported by Macrogen Inc. (grant no. SNU RNDB 0411-20160001 and MGR14-01) and partly supported by Post-Genome Technology Development Program (grant no. 10050164, Developing Korean Reference Genome) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

**Author Contributions** J.-S.S. and C.K. conceived and designed the experiments. J.-Y.S., J.-Y.Y. and Ji.Ki. conducted sequencing and relevant experiments. J.Ku., G.H.P. and Juh.Ki. performed BAC clone library preparation and sequencing. A.H. and H.C. generated BioNano data. H.R. performed RNA-seq and isoform sequencing analysis. Jo.Ki. performed phasing of 10X Genomics linked reads. M.R. and J.B. performed *de novo* assembly. J.Ko. and M.W.H. performed PacBio sequencing. M.-H.S. performed gap closure. Jun.Ki., S.L. and C.-U.K. performed SV analysis. A.R. performed phasing analysis. J.-S.S., A.R., Jun.Ki., S.L., M.-H.S. and C.K. primarily wrote the manuscript, although many authors provided edits.

**Author Information** The accession codes for the underlying sequence data are summarized in Supplementary Table 22. The AK1 assembly and haplotigs have been deposited at DDBJ/ENA/GenBank under the accessions LPV000000000 and LYWJ000000000, respectively. The AK1 assembly, haplotigs, BAC placements and SVs from various platforms and dot plots of gaps resolved by the AK1 assembly are available as a browsable track on AK1 genome browser (<http://211.110.34.36/gbrowse2>). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.-S.S. (jeongsun@snu.ac.kr) or C.K. (kimchan@macrogen.com).

**Reviewer Information** *Nature* thanks S. Salzberg and the other anonymous reviewer(s) for their contribution to the peer review of this work.



This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.



## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

**AK1 cell line.** An immortalized lymphoblastoid cell line was established from the AK1 individual through Epstein–Barr virus transformation of mononuclear cells (Seoul Clinical Laboratories Inc.). Full pathogen testing was performed and maintained in a mycoplasma-free facility. AK1 lymphoblastoid cell line was cultured in RPMI 1640 media containing 15% FBS at 37 °C in a humidified 5% CO<sub>2</sub> environment. The approval number C-0806-023-246 for the AK1 individual was assigned based on the guidelines from the Institutional Review Board of Seoul National University.

**PacBio data generation.** Genomic DNA was extracted from AK1 cells using the Gentra Puregene Cell Kit (Qiagen). Large-insert PacBio library preparation was conducted following the Pacific Biosciences recommended protocols. In brief, a total of 60 µg AK1 genomic DNA was sheared to ~20 kb targeted size by using Covaris g-TUBEs (Covaris). Each shearing processed 10 µg input DNA and a total of 6 shearings were performed. The sheared genomic DNA was examined by Agilent 2100 Bioanalyzer DNA12000 Chip (Agilent Technologies Inc.) for size distribution and underwent DNA damage repair/end repair, blunt-end adaptor ligation followed by exonuclease digestion. The purified digestion products were loaded onto pre-cast 0.6% agarose for 7–50 kb size selection using the BluePippin Size Selection System (Sage Science), and the recovered size-selected library products were purified using 0.5× pre-washed Agencourt AMPure XP beads (Beckman Coulter). The final libraries were examined by Agilent 2100 Bioanalyzer DNA12000 Chip for size distribution and the library concentration was determined with Qubit 2.0 Fluorometer (Life Technologies). We sequenced with the PacBio RSII instrument with P6 polymerase binding and C4 chemistry kits (P6C4). A total of 380 SMRT Cells were used to yield 101-fold whole-genome sequence data.

**Sample preparation for BioNano Genomics.** AK1 cells were pelleted and washed with PBS; the final cell pellet was re-suspended in cell-suspension buffer using the CHEF Mammalian Genomic DNA Plug Kit (Bio-Rad). Cells were then embedded in CleanCut low-melt Agarose (Bio-Rad) and spread into a thin layer on a custom support (in development at BioNano Genomics). Cells were lysed using IrysPrep Lysis Buffer (BioNano Genomics), protease-treated with Puregene Proteinase K (Qiagen), followed by brief washing in Tris with 50 mM EDTA and then washing in Tris with 1 mM EDTA before RNase treatment with Puregene RNase (Qiagen). DNA was then equilibrated in Tris with 50 mM EDTA and incubated overnight at 4 °C before extensive washing in Tris with 0.1 mM EDTA followed by equilibration in NEBuffer 3 (New England BioLabs) at 1× concentration. Purified DNA in the thin layer agarose was labelled following the IrysPrep Reagent Kit protocol with adaptations for labelling in agarose. In brief, 1.25 µg of DNA was digested with 0.7 U Nt.BspQI nicking endonuclease per microlitre of reaction volume in NEBuffer 3 (New England BioLabs) for 130 min at 37 °C, then washed with TE Low EDTA Buffer (Affymetrix), pH 8.0, followed by equilibration with 1× ThermoPol Reaction Buffer (New England BioLabs). Nick-digested DNA was then incubated for 70 min at 50 °C using the IrysPrep Labelling mix (BioNano Genomics) and Taq DNA Polymerase (New England BioLabs) at a final concentration of 0.4 U µl<sup>-1</sup>. Nick-labelled DNA was incubated for 40 min at 37 °C using the IrysPrep Repair mix (BioNano Genomics) and Taq DNA Ligase (New England BioLabs) at a final concentration of 1 U µl<sup>-1</sup>. Labelled-repaired DNA was then recovered from the thin layer agarose by digesting with GELase and counterstained with IrysPrep DNA Stain (BioNano Genomics) before data collection on the Irys System. The fragile site rescue process protects fragile sites by reducing the temperature of the labelling reaction and minimizes shear forces by restraining DNA in agarose until nicks are repaired. In this case, only the closest opposite-strand nick-pairs break.

**Sequencing library preparation using the GemCode platform.** Sample indexing and partition barcoded libraries were prepared using GemCode Gel Bead and Library Kit (10× Genomics)<sup>4</sup>. Sequencing was conducted with Illumina HiSeq2500 to generate linked reads.

**Illumina data generation.** Libraries were generated with PCR-free protocols. gDNA was sheared twice using Covaris S2 with cycling conditions of 10% duty cycle, Cycles/Burst 200, and Time 100 s. The sheared DNA was processed using the Illumina TruSeq DNA PCR-Free LT Library Kit protocol to generate 550 bp inserts, which includes end repair, SPRI bead size selection, A-tailing, and Y-adaptor ligation. Library concentration was measured by qPCR and loaded on HiSeq X Ten instruments (PE-150) to generate 72-fold sequence coverage.

**DNA preparation from BAC clones.** A total of 32,026 BAC clones were selected from the 252 384-well plates and re-plated into 96-well plates. Clones were grown overnight, and the cultures were used to prepare two additional replicates for the two 384-well plates that were stored at –80 °C in LB medium containing 20%

glycerol. A total of 32,026 clone cultures with growth at ODs ranging from 0.6 to 1.0 were pooled, pelleted and the DNA was extracted using the standard alkaline lysis method. In this procedure, a cell pellet was resuspended in 150 µl of Qiagen buffer P1 with RNase and lysed with 150 µl of 0.2 M NaOH, 1% SDS solution for 5 min. Lysis was neutralized with the addition of 150 µl of 3 M sodium acetate, pH 4.8. Neutralized lysate was incubated on ice for 30 min, and DNA was collected by centrifugation for 15 min at 15.7 g at 4 °C, concentrated by standard ethanol precipitation and resuspended in 25 µl of 10 mM Tris-HCl, pH 8.5.

**PacBio sequencing of BAC clones.** DNA from approximately 150 BAC clones with roughly equimolar concentration was combined into a single pool. A total of 10 µg from each pool DNA was sheared and fragments of insert size ranging from 10 to 15 kb were selected. Two libraries were prepared from the pooled DNA using a PacBio SMRTbell library preparation kit v1.0. The libraries were quantified using a Qubit 2.0 fluorometer and each library was sequenced using two SMRT cells with P6C4 chemistry.

**Illumina sequencing of BAC clones.** DNA from approximately 290 BAC clones with roughly equimolar concentration was combined into a single BAC pool. One nanogram of DNA from each pool was digested and fragments of insert size ranging from 500 to 550 bp were selected. In total, 109 libraries were prepared from the pooled DNA using Illumina-compatible Nextera XT DNA sample prep kit and sequenced with HiSeq2500.

**Sample preparation for RNA sequencing.** We extracted RNA from tissue using RNAiso Plus (Takara Bio), followed by purification using RNeasy MinElute (Qiagen). RNA was assessed for quality and was quantified using RNA 6000 Nano LabChip on a 2100 Bioanalyzer (Agilent). The RNA sequencing (RNA-seq) libraries were prepared as previously described<sup>20</sup>. RNA library was sequenced with Illumina TruSeq SBS Kit v3 on a HiSeq2000 sequencer (Illumina) to obtain 100 bp paired-end reads. The image analysis and base calling were performed using the Illumina pipeline with default settings.

**Sample preparation for isoform sequencing.** Total RNA extracted from AK1 cells with RNA integrity number (RIN) > 8.0 was used for library preparation. The library was constructed following the Clontech SMARTer-PCR cDNA Synthesis Sample Preparation Guide. 1–2 kb, 2–3 kb, 3–6 kb and > 5 kb libraries were selected by Sage, ELF purified, end-repaired and blunt-end SMRTbell adapters were ligated. The fragment size distribution was confirmed on a Bioanalyzer HS chip (Agilent) and quantified on a Qubit fluorometer (Life Technologies). The fragment size distribution was validated on a Bioanalyzer HS chip (Agilent) and quantified on a Qubit fluorometer (Life Technologies). The sequencing was carried out on the PacBio RSII instrument using P6C4.

**PacBio long-read *de novo* assembly.** Around 31 million subreads were used for assembly with FALCON v0.3.0 (ref. 21) given length\_cutoff parameter of 10 kb for initial mapping to build pre-assembled reads (preads), and preads over 15 kb were used (length\_cutoff\_pr) to maximize the assembled contig N50 (Extended Data Fig. 2). Primary and associated contigs were polished using Quiver<sup>5</sup>.

**BioNano Genomics genome map generation.** Optical maps were *de novo* assembled into genome maps using BioNano assembler software (Irys System, BioNano Genomics). Single molecules longer than 150 kb with at least 8 fluorescent labels were used to find possible overlaps ( $P < 1 \times 10^{-10}$ ). Next, these maps were constructed to consensus maps by recursively refining and extending them by mapping single molecules ( $P < 1 \times 10^{-5}$ ). The consensus maps were compared and merged into genome maps when patterns matched ( $P < 1 \times 10^{-10}$ ). A second set of optical maps was obtained thereafter, and generated into genome maps with the same criteria.

**Contig editing and hybrid assembly.** Primary contigs were *in silico* digested into cmaps and were compared with genome maps for scaffolding. The scaffolding was visualized and performed with the Irys Viewer. When conflict occurred, the contigs were edited with the guidance of genome map.

**Assembly improvements.** Paired-end reads from Illumina platform were aligned to the assembly using bwa<sup>22</sup> mem, followed with duplication removal using Picard tools<sup>23</sup>. Base-pair correction of the assembly was performed using Pilon<sup>24</sup>. Pilon mostly corrected single insertions and deletions in regions enriched with homopolymer. Contigs or scaffolds shorter than 10 kb were excluded from the overall analysis to avoid results from spurious misassembly.

**Scaffold accuracy measurement with BAC clones.** Scaffolding accuracy of the AK1 assembly was assessed using the AK1 BAC library<sup>1</sup>. AK1 BAC end sequences (BES) were aligned to GRCh37, GRCh38 and AK1 assemblies using BWA. The BES placements were categorized by the alignment, orientation and separation of BES with respect to the assembly. The BES placement was determined to be concordant: (1) if the BES placement was placed in the same assembly unit; (2) if the paired end sequences were properly oriented; and (3) if the *in silico* insert size was between 50,000 and 250,000 bp. If the BES placements did not meet these conditions, the BES placement was defined to be discordant. In addition, if only one

of the paired end sequences were aligned to the assembly, the BES placement was defined to be an orphan placement. If both paired-end sequences were unaligned to the assembly, the BES was defined to be unmapped. If either of the paired-end sequences were aligned to different positions of the assembly multiple times, the BES was defined to have multiple placements.

**Gap closure and SV analysis: alignment to the reference genome.** To identify the precise genomic location of each assembly unit, we used LASTZ<sup>25</sup> with parameters (-gapped -gap = 600,150,-hspthresh = 4500,-seed = 12of19 -notransition -ydrop = 15000-chain) to align each assembly unit to each chromosome in the human reference genome. Chaining procedure was followed to join the neighbouring local alignments into a single cohesive alignment. The chained alignments of each assembly unit were processed to obtain a single alignment with the best alignment score. If the selected alignment was not fully representative of the assembly unit, we selected a set of alignments that was better representative of the assembly unit. A netting procedure was then followed with the selected chained alignments. The chaining and netting procedures were applied using UCSC Kent tools<sup>26</sup> and parallel processing was used when possible to increase computational speed.

**Gap closure of GRCh38.** Gaps were classified into telomeric, centromeric, heterochromatic, acrocentric and euchromatic region according to the agp file and cytoband information provided by the Genome Reference Consortium (GRC) and UCSC genome browser. In total, 190 euchromatic gaps were targeted for gap closure with AK1 assembly. The gaps that could not be closed or extended with the AK1 assembly were subjected to closure through local assembly using Canu<sup>27</sup> or a contiguous subread. Subreads mapped 10 kb upstream or downstream of the gap were chosen for local assembly. Alignment was performed with BLASR<sup>28</sup> -bestn 3, and primary aligned reads with mapping quality of 254 were used. The assembled contigs were thereafter aligned to their respective gap position to precisely identify the added sequences. Subreads used to close the gaps were chosen following criteria described in the Supplementary Information.

**Assembly based variant detection.** The alignments of the assembly to the reference genome were parsed to obtain SNPs, indels and SVs, which we defined as insertion, deletion, inversion and complex variants with event size equal to or greater than 50 bp. The complex SVs are the same as 'double-sided insertion' defined previously<sup>29</sup>. We used GRCh37 instead of GRCh38 for the main analysis for compatibility and comparison with previously reported structural variations.

**SV annotation.** Repeat elements were annotated using RepeatMasker (-species human -no\_is) and tandem repeat finder (TRF) (2 7 7 80 10 50 2000 -f -m -h -d). SVs are classified accordingly if it is masked by at least 70% with a single type. Complex is defined as the SVs having either several annotated repeat elements, or at least 30% of the remaining sequence not annotated as repeat. Novelty was identified by comparing the breakpoints with 50% reciprocal overlap criterion. Functional annotation was performed using both GENCODE release v19 (GRCh37) and v21 (GRCh38)<sup>30</sup> and the Ensembl Regulatory Build<sup>31</sup>. For those SVs that occurred within gene regulatory domains, we annotated with the nearest gene name. SV located within pericentromeric regions (5 Mb flanking annotated centromeres) and subtelomeric regions (150 kb from the annotated telomeric sequence) were annotated as heterochromatin. Both pilot and strict accessibility genome mask regions (version 20141020) were downloaded from [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible\\_genome\\_masks/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/). Segmental duplication sites were downloaded from the UCSC table browser. To simplify categorization of the SVs that lie within multiple functional regions, they were classified according to the order of priority as follow: coding sequence, untranslated region, intron, transcription-factor-binding site, promoter, enhancer, CTCF (transcriptional repressor), and intergenic. To annotate whether the SVs called from GRCh37 were also shared with GRCh38 SV sets, we compared each AK1 breakpoints with 50% reciprocal overlap criterion. In addition, we assessed whether the SVs called from GRCh38 were also represented in the alternative contigs by measuring the concordance against the SV regions including the surrounding 50 bp from the breakpoints.

**Asian-specific SVs.** Population allele frequency of SVs was obtained by aligning reads from 38 high-coverage samples from five different ancestral backgrounds (African, American, European, East Asian, and South Asian) to the AK1 assembly. We obtained whole-genome sequencing data of 23 individuals from the 1000 Genomes Project and we additionally sequenced 15 East Asian individuals (5 Japanese, 5 Chinese and 5 Koreans). Analysis candidates were selected from the insertions with less than 70% of repeats. We excluded any duplications among the insertions that are mapped to GRCh37 using BLAST (-evalue 1e-10 -perc\_identity 90 -qcov\_hsp\_perc 90). The regions that have been recognized as mobile element or tandem repeat by RepeatMasker and TRF softwares were masked for analysis. Normalized read depth within the unique sequence was achieved by dividing the read depth, which was calculated using samtools bedcov, by the

median genome coverage. The insertions were determined to be highly polymorphic if there were greater than or equal to 0.3 variant frequency differences across the different populations. Asian-specific insertions were chosen by selecting the insertions with equal or above 0.3 allele frequency difference between Asian and non-Asian population as well as non-Asian allele frequency with equal or below 0.5. Asian linkage disequilibrium blocks were obtained from East Asian samples in the 1,000 Genomes Project phase 3 using S-MIG++ algorithm<sup>32</sup> (-maf 0.05 -ci AV -probability 0.95). Linkage disequilibrium blocks with below 0.8 haplotype diversity index were excluded.

**De novo phasing markers.** We performed phasing against the *de novo* assembly. SNPs and short indels called from whole-genome sequencing (72×) of short reads were phased with linked reads. The non-redundant set of PacBio subreads were aligned to the assembly, and corrections were applied by calculating the maximum likely variant allele for the phased variants based on the read depth. A phased block was defined as the region spanning two markers which had a subread or linked read information providing phasing. Similar to the linked reads, Illumina sequenced BAC phase information was used to correct phasing markers and extend phased blocks. Correction and other bioinformatics methods were performed using an in-house script, described in the Supplementary Information.

**Switch error of phasing markers.** Long-range switch error measurements were obtained using BAC end sequences. The end sequences were aligned to the AK1 assembly with bwa mem, and the base allele of the phasing marker site was called with the corresponding BAC information. When switching occurred for more than two marker sites in a phased block, it was defined as a long range switch. The long-range switch error rate was calculated as: no. of long range switches/no. of phasing markers.

**Haplotig assembly.** Using the final set of phasing markers, subreads were classified into sets of haplotype A or B when >85% of the phasing markers agreed. When a subread contained no marker, it was classified as homozygous. Through the read depth, phasing markers that were missed in previous steps were additionally called for homozygous regions adjacent to phased blocks. Subreads in haplotype A or homozygous regions were assembled into haplotig A, and haplotype B into haplotig B with Canu<sup>27</sup>. Haplotigs for MHC class I and II were assembled separately to avoid misassemblies owing to high sequence homology between HLA genes. In this case, subreads phased as homozygous were used with subreads of haplotype A and B. Homozygously phased subreads flanked on each side of a sequencing gap belonged on haplotype A and B, respectively, and were re-classified during assembly.

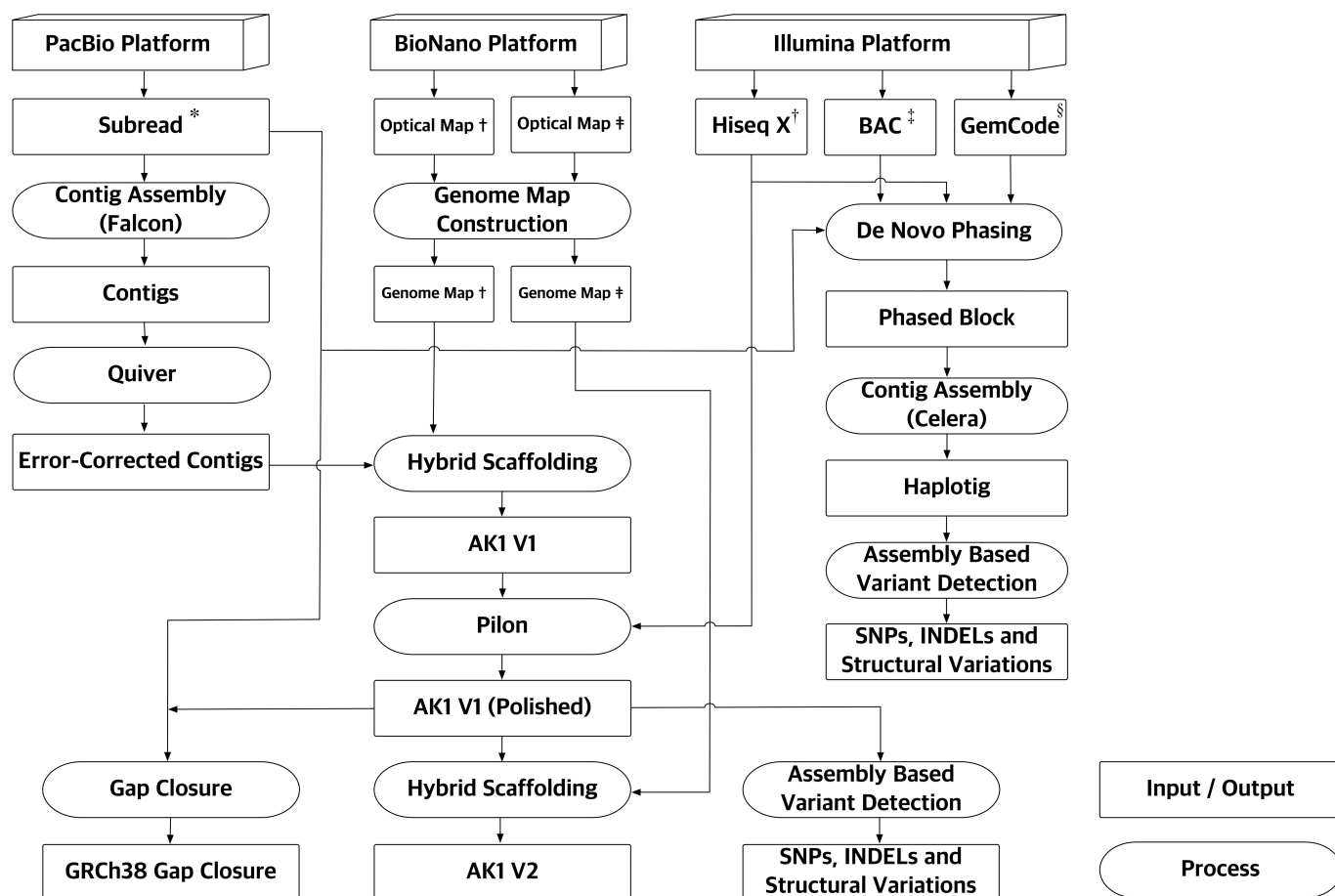
**Haplotype-specific variant calling and annotation.** Haplotype-specific variants were called following the assembly-based variation calling method. Owing to possibilities of false positives introduced by misassembly, phased variants that agreed with initial variants called with whole genome sequencing reads were used for further analysis. After functional annotation using GENCODE v19 (ref. 30), disease risk alleles were screened using ClinVar<sup>33</sup>. Haplotyping of CYP2D6 was done by comparing haplotigs to M33388 following CYP2D6 nomenclature.

**De novo assembly of BAC clones and SV validation.** BACs identified to be discordant in size (>1 kb) were pooled and sequenced with the SMRT platform. The subreads were assembled using Canu<sup>27</sup> after screening and removing *Escherichia coli* or vector sequences with CrossMatch<sup>34</sup>. The assembled BAC contigs were polished with Quiver. The BAC contigs were, thereafter, used to validate AK1 assembly-based or phase-specific SVs by assessing the concordance between the assembly and the BAC contig at sites of detected SVs.

**Heterozygosity and allele-specific expression.** On the basis of the alignments of haplotigs to GRCh37, haplotig A and B were localized to compare partner sequences. The number of different bases were summed in every 5 Mb distance, and percentiled to draw in the Fig. 3a. RNA-seq reads were trimmed and aligned to GRCh37 using STAR aligner<sup>35</sup> with the two-pass mapping strategy as recommended. Duplicates were removed using Picard tools, and variants were called using HaplotypeCaller and VariantFiltration following GATK best practices on RNA-seq<sup>36</sup>. Sites with supportive evidence of altered variation in RNA-seq have been extracted from the final vcf file, and ASEReadCounter<sup>37</sup> was applied to remove reads with low base quality. Read counts are annotated to the phase-specific variants called from haplotigs using in-house scripts. When read depth for one allele was over 30, it was considered as 'expressed'.

20. Seo, J.-S. *et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Res.* **22**, 2109–2119 (2012).
21. PacificBiosciences/FALCON; GitHub, available at: <https://github.com/PacificBiosciences/FALCON> (accessed 2 August 2016)
22. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
23. Picard Tools by Broad Institute; available at: <http://broadinstitute.github.io/picard> (accessed 2 August 2016)

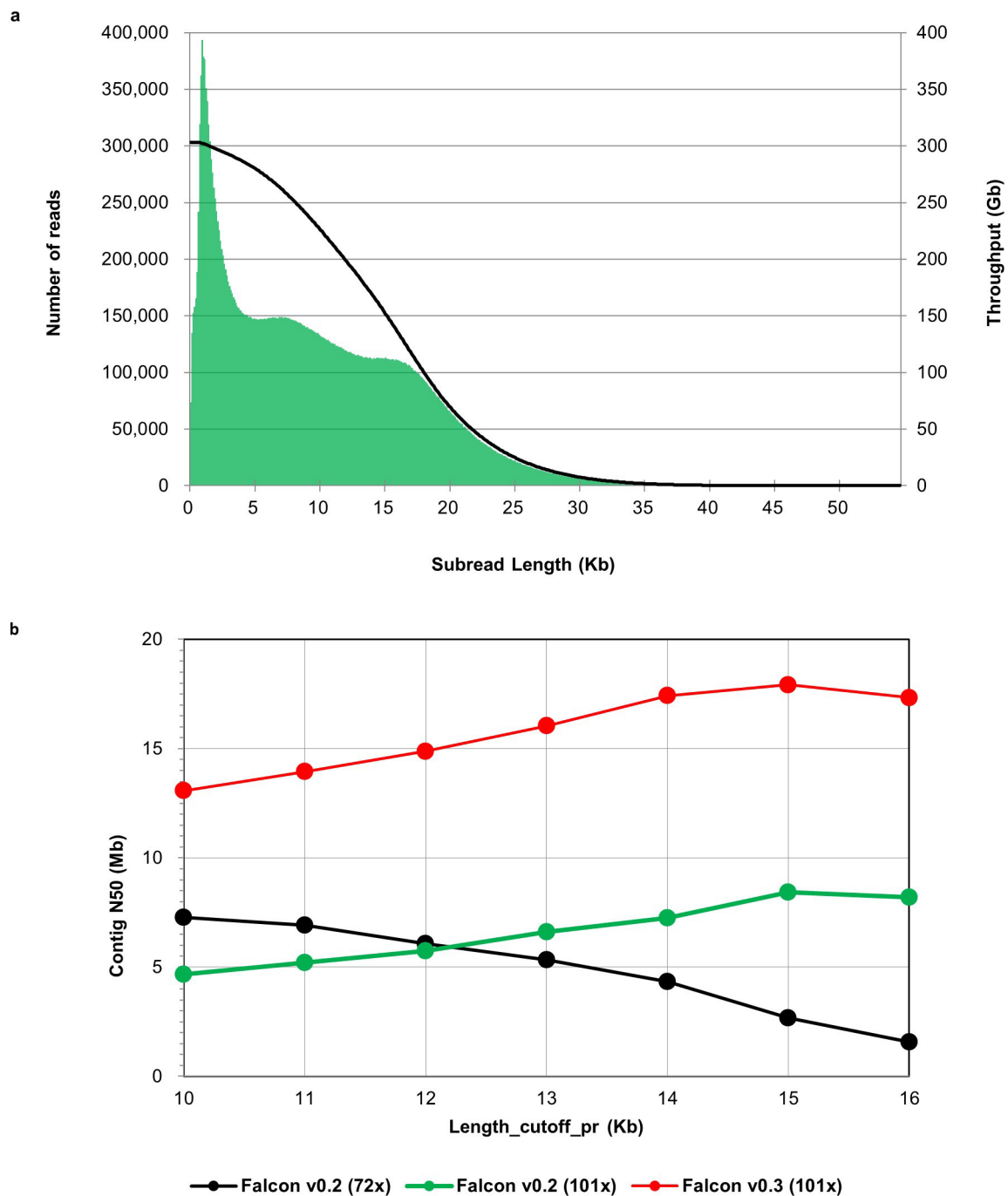
24. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
25. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, Penn State Univ. (2007).
26. ENCODE-DCC/kentUtils; GitHub, available at: <https://github.com/ENCODE-DCC/kentUtils> (accessed 2 August 2016)
27. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R. & Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Preprint at <http://dx.doi.org/10.1101/071282> (2016).
28. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
29. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100**, 11484–11489 (2003).
30. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
31. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).
32. Taliun, D., Gamper, J., Leser, U. & Pattaro, C. Fast sampling-based whole-genome haplotype block recognition. *IEEE/AMC Trans. Comput. Biol. Bioinf.* **13**, 315–325 (2016).
33. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
34. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
35. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
36. The GATK Best Practices for variant calling on RNAseq, in full detail; available at: <http://gatkforums.broadinstitute.org/wdl/discussion/3892/the-gatk-best-practices-for-variant-calling-on-rnaseq-in-full-detail> (accessed 2 August 2016).
37. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).



**Extended Data Figure 1 | Global overview of data generation and sequencing throughput.** Flowchart of the data generation, processing and analysing for the *de novo* assembly and haplotype phasing of the AK1 diploid genome. \*The SMRT platform sequencing throughput is described in Supplementary Table 1. †The number of read and sequencing throughputs from the Illumina platform are 1,635,192,864

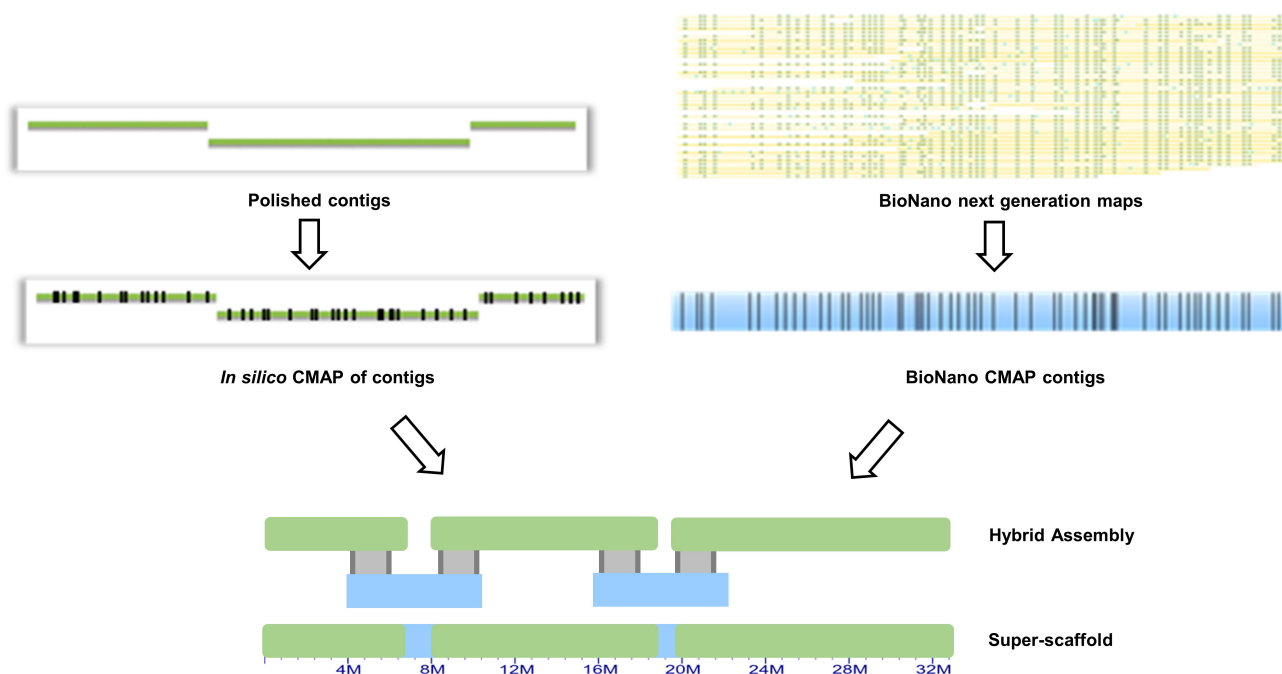
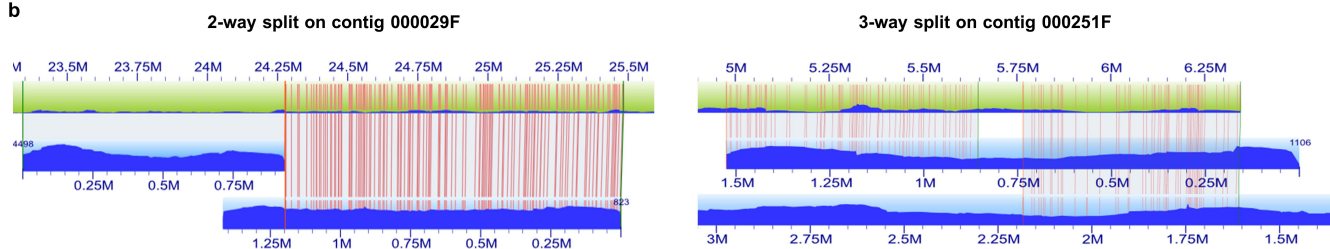
and 249,914,122,464 bp, respectively. ‡AK1 BAC library was sequenced using Sanger capillary end sequencing (single end: 22,563, paired end: 62,758), Illumina (31,719) and SMRT (307) platform. §Linked-read data were additionally generated with the GemCode platform to produce 1,153,598,732 reads from high molecular mass DNA with an average insert size of 100 kb.





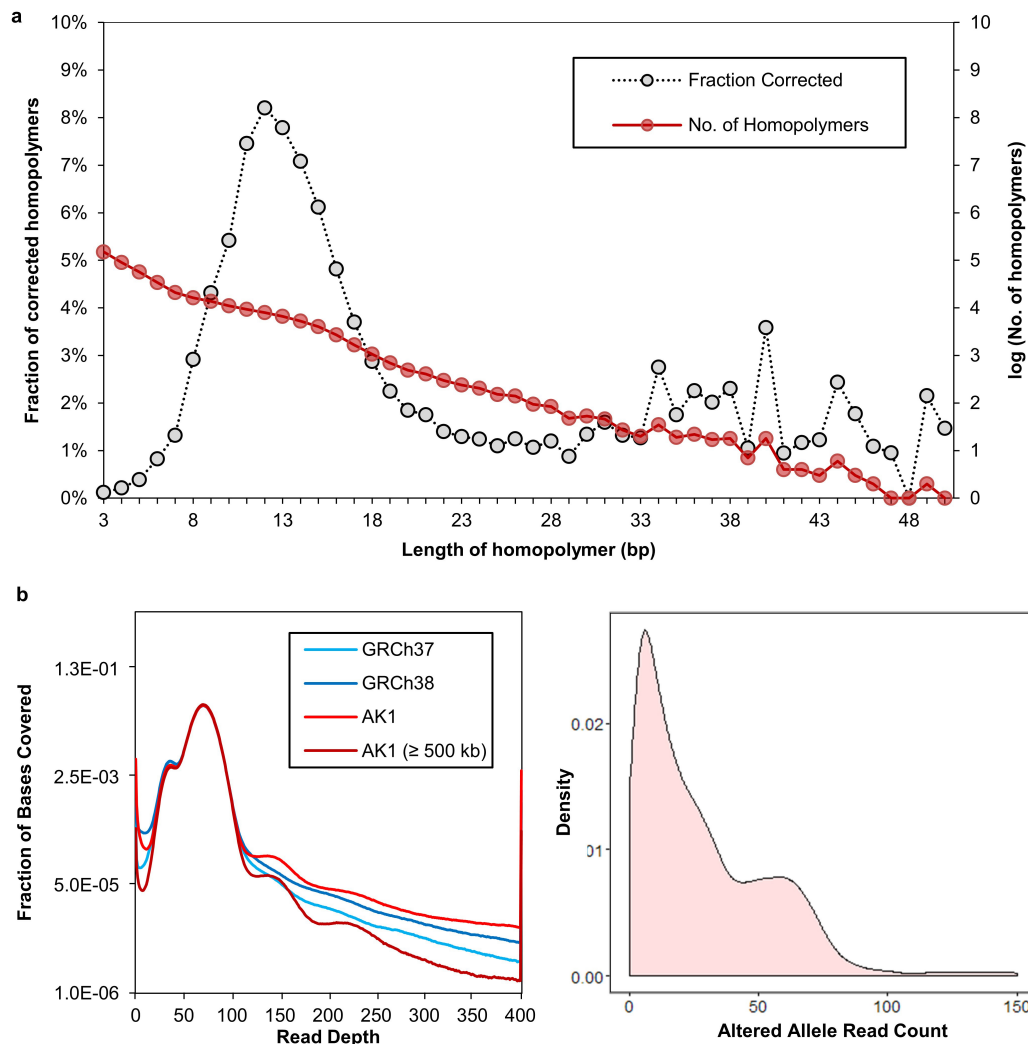
**Extended Data Figure 2 | Length distribution of SMRT subreads and FALCON parameter optimization for assembly.** **a**, The y axis on the left shows the number of subreads with given length (bin size = 100 bp) on the x axis, whereas the y axis on the right shows the sum of the length of subreads longer than or equal to the given length on the x axis. **b**, Effects of length cutoff parameters on contig N50 in *de novo* assembly by FALCON

is shown on the right. The contig N50 depends on the two parameters, related to the amount of error-corrected reads for final assembly, length\_cutoff and length\_cutoff\_pr, respectively, where the former was fixed at 10 kb but the latter varied from 10 to 16 kb. Black and green lines indicate the changes of N50 for 72× and 101× sequencing dataset, respectively.

**a****b**

**Extended Data Figure 3 | Graphical representation of hybrid assembly and statistics for next generation map and genome map. a,** The hybrid assembly approach aligns *in silico* generated maps from sequence contigs with genome maps. When genome maps bridge two contigs, a scaffold is produced. The comparison is visualized between the genome maps and

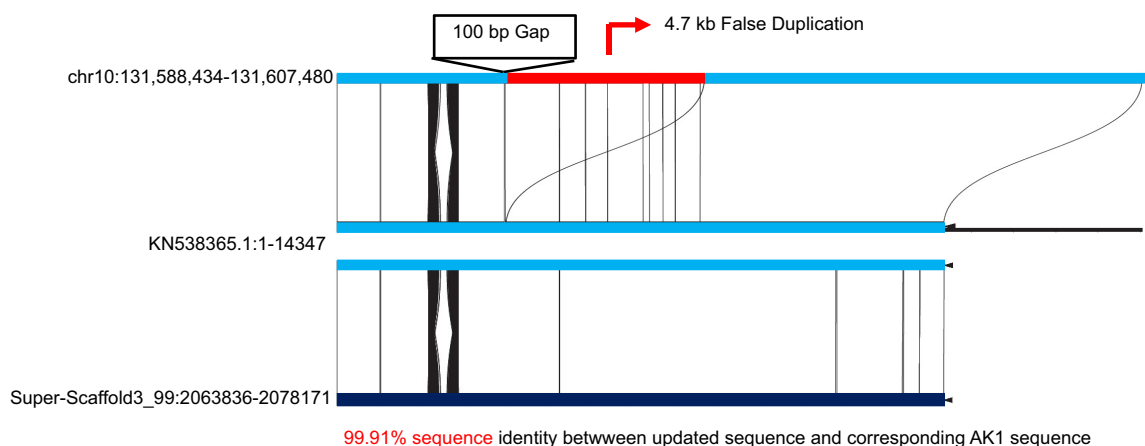
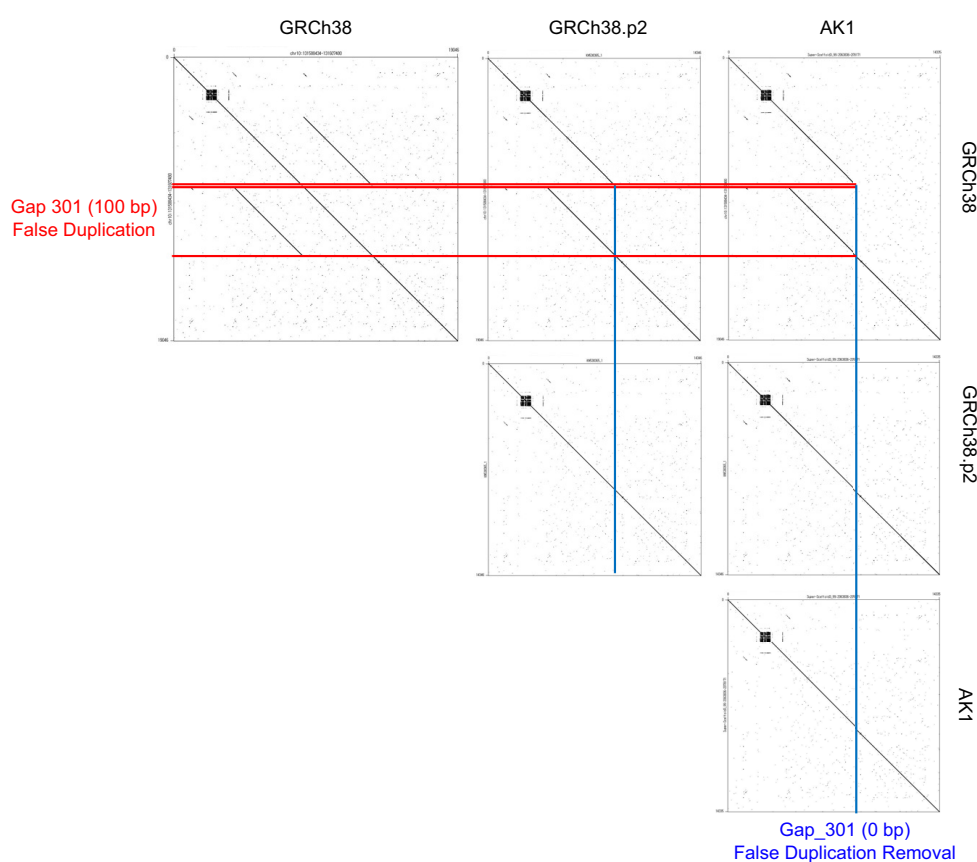
contigs in the Iris Viewer. **b,** Examples of edited contigs due to conflicts between the contig and the genome maps. The matches between the *in silico* map and the genome map are highlighted in red, and mismatches are indicated by absence of the red lines.



**Extended Data Figure 4 | Assessment of assembly accuracy with homopolymer and read depth coverage generated with short reads.**

**a**, Distribution of corrections in homopolymer. Pilon mostly corrected the single base deletions in the assembly and the corrections are enriched in regions with long stretches of homopolymer. **b**, The read-depth distribution against AK1 assembly, AK1 assembly with scaffolds  $\geq 500$  kb, GRCh37 and GRCh38. As the mean coverage depth of short reads was  $72\times$ , a peak is shown around it representing the fraction of autosomal region. Another peak is shown in  $\sim 36\times$ , which is half of the mean

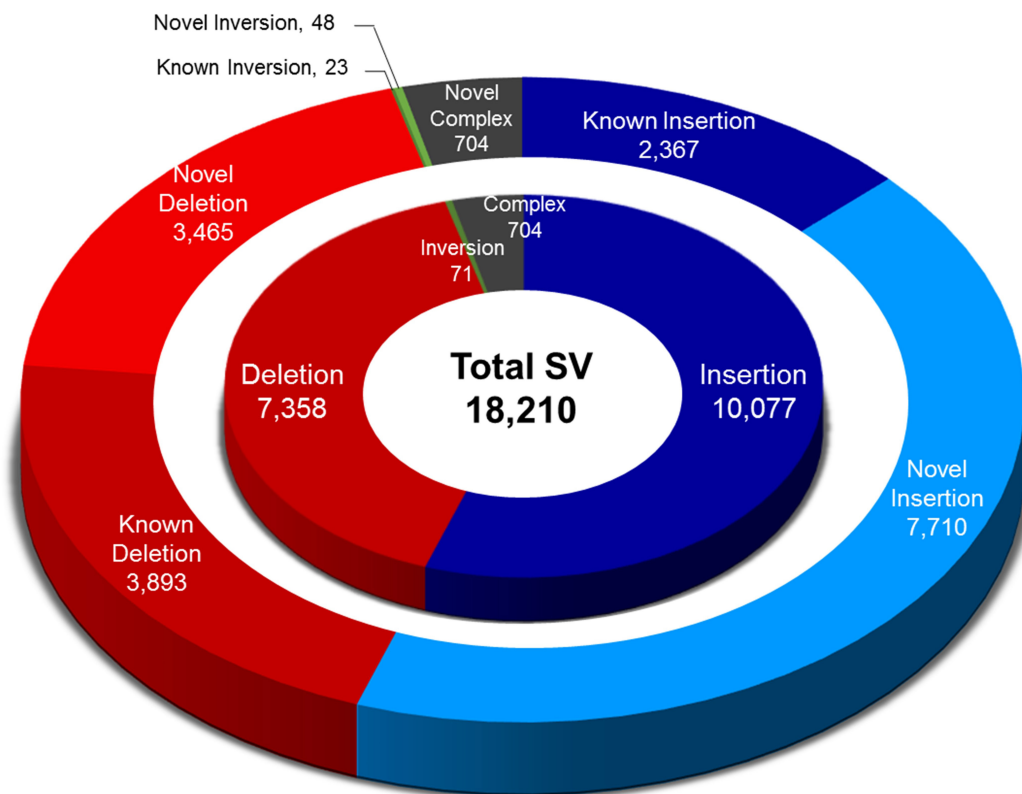
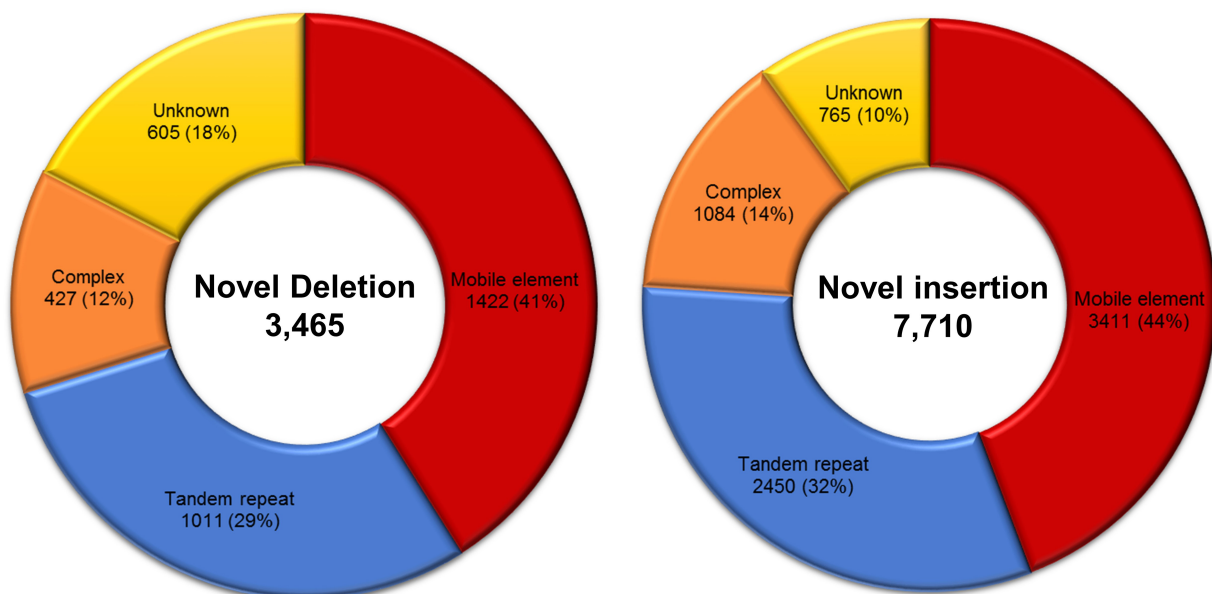
coverage depth, representing the contigs derived from chromosomes X and Y. The fluctuating long tale is showing 3-copy and 4-copy of a haplotype, but more clearly observed with AK1 long scaffolds. The overall pattern is showing that more SVs are reflected in AK1 long contigs than the reference. The short contigs ( $< 500$  kb) are only 120.4 Mb, comprising a small fraction of the AK1 assembly. **c**, Density plot of the homozygously altered allele read depth from long scaffolds ( $\geq 500$  kb). Most variants are skewed in low allelic read depth, suggested to be mainly due to sequencing artefact or mapping bias.

**a****b**

**Extended Data Figure 5 | An example of filled sequence that matches perfectly with the patch sequence (KN538365.1).** **a**, One AK1 scaffold (Super-scaffold3\_99) closes a 100-bp gap in chromosome 10, reducing the size of this gap to zero while it also removes a 4.7 kb false duplication found left of the gap. This information corresponds perfectly to the GRCh38 fix patch (KN538365.1) sequence covering this region, thus

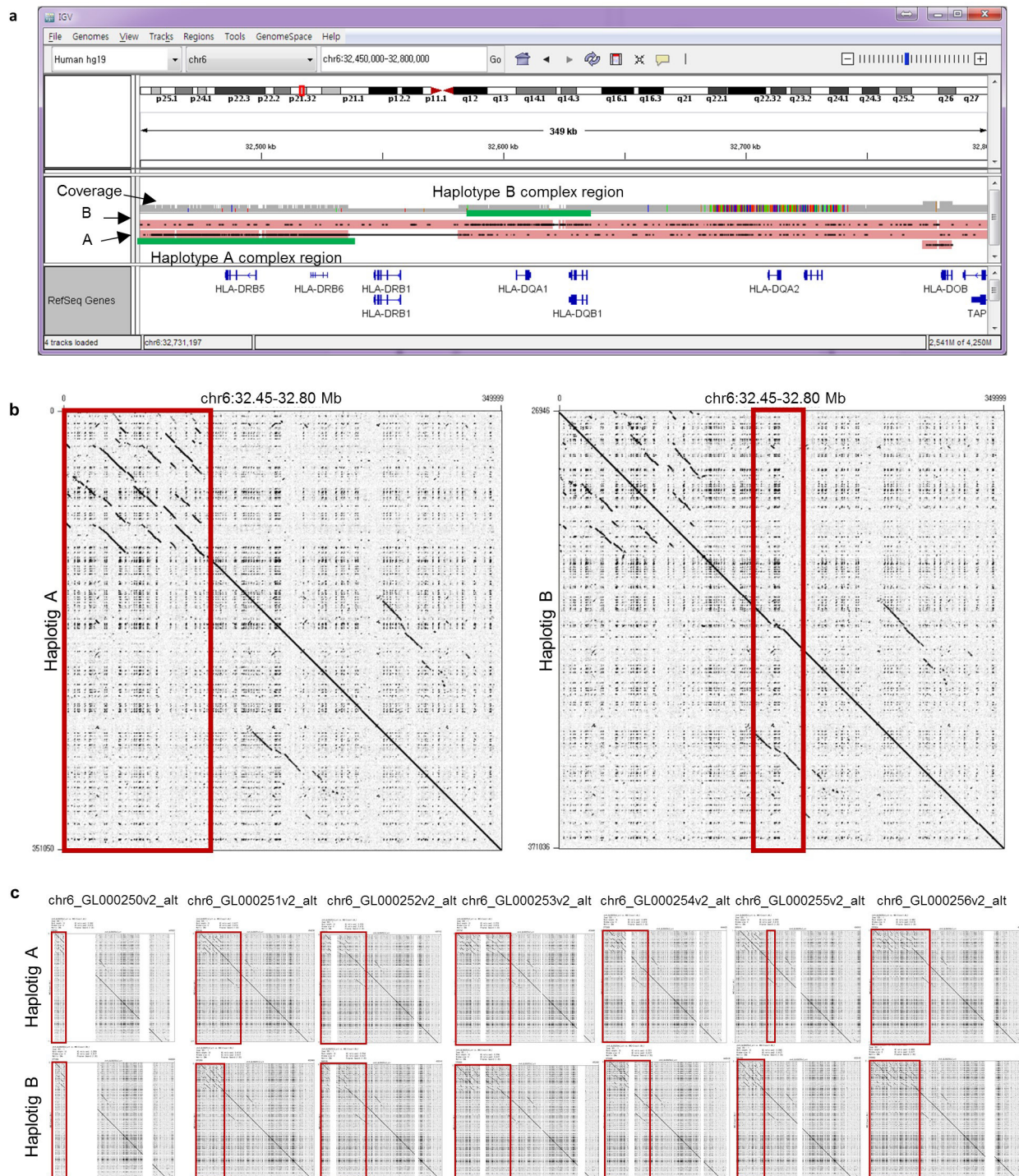
validating our assembly and gap closing accuracy. **b**, Six dot plots show the comparison between GRCh38, KN538365.1 and the AK1 assembly. The dot plots are organized in the following manner: Reference-reference (top left), KN538365.1-reference (top middle), AK1-reference (top right), KN538365.1-KN538365.1 (centre middle), AK1-KN538365.1 (middle right) and AK1-AK1 (bottom right).



**a****b****Extended Data Figure 6 | Number of SVs and repeat composition.**

**a**, Overall distribution of SVs. By direct comparison between AK1 assembly and GRCh37 reference genome, deletion (red), insertion (blue), inversion (green), and complex (grey) variants were detected. Outer pie chart represents new variants for each SV types. In total, 65% (11,927)

of the SVs were unreported previously. **b**, Repeat composition of AK1 insertion and deletion. Both insertions and deletions are mostly composed of mobile elements or tandem repeats. Complex is defined as the SVs having either several annotated repeat elements, or at least 30% of the remaining sequence not annotated as repeat.



**Extended Data Figure 7 | MHC class II haplotigs alignment on chromosome 6 and dot plots. a,** MHC haplotigs A and B aligned on GRCh37 chr6. The complex regions shown in Fig. 3a are in green bars. **b,** Dot plot of haplotig A and B to the reference genome. The region highlighted in red is giving many SVs when aligning on the reference

owing to different sequence context in haplotigs. **c,** Dot plots of haplotig A and B to the alternative loci (ALT) patches of MHC region in hg38. Haplotype A had the most similarities with chr6\_GL000255v2\_alt for the highlighted region in b. The blank vertical lines indicate 'N' bases in the reference ALT sequence.

Extended Data Table 1 | Summary of *de novo* assembly and phasing statistics

Data set		No. of contigs or scaffolds	Sum (Gb)	N50 (Mb)	Longest (Mb)	Average (Mb)	L50	No. of gaps	Percent bases of gaps
Contigs	Assembly	3,128	2.87	17.7	76.5	0.92	50	0	0.00
Scaffold V1	Scaffolding	2,927	2.89	29.1	113.9	0.99	28	188	0.75
Scaffold V2	Scaffolding with Fsr-GM	2,832	2.90	44.8	113.9	1.03	21	264	1.29
Phased Block	Linked-reads	1,468	2.62	5.70	29.6	1.78	143	n/a	
	Linked-reads with BAC	836	2.64	11.55	65.0	3.16	71		
Haplotig	A	3,155	2.63	2.41	11.21	0.83	328	n/a	
	B	15,816	2.19	0.32	2.52	0.14	2,012		

Phased blocks are measured by sub-read coverage based on markers on each step. These subreads are locally assembled to build haplotigs. A, haplotig assembled with subreads phased as haplotype A and homozygous; B, haplotig B assembled with subreads phased as haplotype B; BAC, bacterial artificial chromosome sequenced with short reads; Fsr-GM, fragile site rescued Genome Map; n/a, not applicable.

Extended Data Table 2 | BAC clone paired-end sequence placements to AK1 and human reference

Type of Placement	GRCh37	GRCh38	AK1 v1	AK1 v2
Paired End Placements	62,076	62,091	62,038	62,022
Unique Placements	60,585	60,460	61,152	61,132
Concordant Placements	56,328	56,486	58,359	58,340
Scaffolding Placements	0	0	3,027	2,928
Discordant Placements	4,257	3,974	2,793	2,792
Discordant in Size	1,481	1,358	1,310	1,316
Discordant in Orientation	1,619	1,555	1,483	1,476
Discordant in Chromosomes	1,157	1,061	na	na
Multiple Placements	1,491	1,631	886	890
Orphan Placements	655	642	698	713
Unmapped	27	25	22	23
Mean in silico insert size (bp)	102,788	102,785	102,928	102,947
In silico insert size standard deviation (bp)	23,149	23,151	23,263	23,295
Total	62,758	62,758	62,758	62,758

The summary of BAC paired end-read alignments to the human reference and the AK1 assembly. The end sequence placements indicate that the assembly quality of AK1 V2 is comparable to that of the human reference genome.



Extended Data Table 3 | Haplotig length and reference coverage on each chromosome

	Haplotig A					Haplotig B				
	No. haplotigs	Bases (Mb)	Coverage	N50 (Kb)	Longest (Mb)	No. haplotigs	Bases (Mb)	Coverage	N50 (Kb)	Longest (Mb)
chr1	175	217.1	96%	2,218.1	7.8	1,262	182.1	81%	337.8	1.5
chr2	179	229.9	96%	2,325.1	9.3	1,403	193.8	81%	315.5	1.5
chr3	139	194.8	100%	2,282.2	8.8	1,250	156.7	80%	297.8	1.6
chr4	140	174.6	93%	2,380.3	10.7	1,084	153.0	82%	311.4	1.6
chr5	122	172.9	97%	2,296.1	11.3	1,064	143.2	81%	300.3	1.3
chr6	104	165.7	99%	2,048.4	6.7	969	136.5	82%	325.0	1.4
chr7	123	149.1	96%	2,294.4	11.8	878	125.3	81%	316.0	1.8
chr8	97	144.2	101%	2,340.6	7.4	850	121.6	85%	316.0	3.4
chr9	88	112.1	93%	2,746.8	7.2	565	96.6	80%	377.8	2.6
chr10	114	125.1	95%	2,485.2	4.5	770	106.1	81%	328.2	1.4
chr11	100	126.1	96%	2,606.4	7.9	724	106.8	81%	346.5	1.8
chr12	73	128.5	98%	2,958.9	10.9	717	107.5	82%	312.1	1.5
chr13	56	95.0	99%	2,975.3	7.4	572	79.1	83%	286.6	1.2
chr14	56	86.1	97%	2,675.1	6.6	457	75.2	85%	361.8	1.5
chr15	74	76.5	94%	2,530.2	6.1	431	63.5	78%	304.8	1.0
chr16	92	73.0	93%	2,196.8	7.3	433	61.7	78%	305.7	1.2
chr17	70	74.8	96%	1,367.1	5.4	489	61.8	79%	269.0	1.3
chr18	43	73.6	99%	3,309.4	7.7	414	64.1	86%	296.5	1.3
chr19	51	55.0	99%	1,853.4	4.4	316	48.2	86%	291.6	1.6
chr20	31	57.9	97%	3,914.5	7.6	328	49.8	84%	349.4	1.5
chr21	35	33.0	94%	1,980.9	12.5	180	29.1	83%	304.1	0.8
chr22	30	33.1	95%	2,429.5	6.2	159	29.5	85%	366.0	1.4
Total	1,992	2,598.1	97%	2,403.9	12.5	15,315	2,191.5	82%	318.7	3.4

Coverage was calculated over autosomal non-N base length of the reference (GRCh37). Haplotig A in chromosome 8 is even longer than the non-N bases of the reference, indicating both assembly and phasing achieved higher contiguity.

# Genome-wide associations for birth weight and correlations with adult disease

Momoko Horikoshi<sup>1,2\*</sup>, Robin N. Beaumont<sup>3\*</sup>, Felix R. Day<sup>4\*</sup>, Nicole M. Warrington<sup>5,6\*</sup>, Marjolein N. Kooijman<sup>7,8,9\*</sup>, Juan Fernandez-Tajes<sup>1\*</sup>, Bjarke Feenstra<sup>10</sup>, Natalie R. van Zuydam<sup>1,2</sup>, Kyle J. Gaulton<sup>1,11</sup>, Niels Grarup<sup>12</sup>, Jonathan P. Bradfield<sup>13</sup>, David P. Strachan<sup>14</sup>, Ruifang Li-Gao<sup>15</sup>, Tarunveer S. Ahluwalia<sup>12,16,17</sup>, Eskil Kreiner<sup>16</sup>, Rico Rueedi<sup>18,19</sup>, Leo-Pekka Lyytikäinen<sup>20,21</sup>, Diana L. Cousminer<sup>22,23,24</sup>, Ying Wu<sup>25</sup>, Elisabeth Thiering<sup>26,27</sup>, Carol A. Wang<sup>6</sup>, Christian T. Have<sup>12</sup>, Jouke-Jan Hottenga<sup>28</sup>, Natalia Vilor-Tejedor<sup>29,30,31</sup>, Peter K. Joshi<sup>32</sup>, Eileen Tai Hui Boh<sup>33</sup>, Ioanna Ntalla<sup>34,35</sup>, Niina Pitkänen<sup>36</sup>, Anubha Mahajan<sup>1</sup>, Elisabeth M. van Leeuwen<sup>8</sup>, Raimo Joro<sup>37</sup>, Vasiliki Lagou<sup>1,38,39</sup>, Michael Nodzenski<sup>40</sup>, Louise A. Diver<sup>41</sup>, Krina T. Zondervan<sup>1,42</sup>, Mariona Bustamante<sup>29,30,31,43</sup>, Pedro Marques-Vidal<sup>44</sup>, Josep M. Mercader<sup>45</sup>, Amanda J. Bennett<sup>2</sup>, Nilufer Rahmioglu<sup>1</sup>, Dale R. Nyholt<sup>46</sup>, Ronald C. W. Ma<sup>47,48,49</sup>, Claudia H. T. Tam<sup>47</sup>, Wing Hung Tam<sup>50</sup>, CHARGE Consortium Hematology Working Group†, Santhi K. Ganesh<sup>51</sup>, Frank J. A. van Rooij<sup>8</sup>, Samuel E. Jones<sup>3</sup>, Po-Ru Loh<sup>52,53</sup>, Katherine S. Ruth<sup>3</sup>, Marcus A. Tuke<sup>3</sup>, Jessica Tyrrell<sup>3,54</sup>, Andrew R. Wood<sup>3</sup>, Hanieh Yaghootkar<sup>3</sup>, Denise M. Scholtens<sup>40</sup>, Lavinia Paternoster<sup>55,56</sup>, Inga Prokopenko<sup>1,57</sup>, Peter Kovacs<sup>58</sup>, Mustafa Atalay<sup>37</sup>, Sara M. Willems<sup>8</sup>, Kalliope Panoutsopoulou<sup>59</sup>, Xu Wang<sup>33</sup>, Lisbeth Carstensen<sup>10</sup>, Frank Geller<sup>10</sup>, Katharina E. Schraut<sup>32</sup>, Mario Murcia<sup>31,60</sup>, Catharina E. M. van Beijsterveldt<sup>28</sup>, Gonneke Willemsen<sup>28</sup>, Emil V. R. Appel<sup>12</sup>, Cilius E. Fonvig<sup>12,61</sup>, Caecilie Trier<sup>12,61</sup>, Carla M. T. Tiesler<sup>26,27</sup>, Marie Standl<sup>26</sup>, Zoltán Kutalik<sup>19,62</sup>, Sílvia Bonàs-Guarch<sup>45</sup>, David M. Hougaard<sup>63,64</sup>, Friman Sánchez<sup>45,65</sup>, David Torrents<sup>45,66</sup>, Johannes Waage<sup>16</sup>, Mads V. Hollegaard<sup>63,64,‡</sup>, Hugoline G. de Haan<sup>15</sup>, Frits R. Rosendaal<sup>15</sup>, Carolina Medina-Gomez<sup>7,8,67</sup>, Susan M. Ring<sup>55,56</sup>, Gibran Hemani<sup>55,56</sup>, George McMahon<sup>56</sup>, Neil R. Robertson<sup>1,2</sup>, Christopher J. Groves<sup>2</sup>, Claudia Langenberg<sup>4</sup>, Jian'an Luan<sup>4</sup>, Robert A. Scott<sup>4</sup>, Jing Hua Zhao<sup>4</sup>, Frank D. Mentch<sup>13</sup>, Scott M. MacKenzie<sup>41</sup>, Rebecca M. Reynolds<sup>68</sup>, Early Growth Genetics (EGG) Consortium†, William L. Lowe Jr<sup>69</sup>, Anke Tönjes<sup>70</sup>, Michael Stumvoll<sup>58,70</sup>, Virpi Lindi<sup>37</sup>, Timo A. Lakka<sup>37,71,72</sup>, Cornelia M. van Duijn<sup>8</sup>, Wieland Kiess<sup>73</sup>, Antje Körner<sup>58,73</sup>, Thorkild I. A. Sørensen<sup>55,56,74,75</sup>, Harri Niinikoski<sup>76,77</sup>, Katja Pakkala<sup>36,78</sup>, Olli T. Raitakari<sup>36,79</sup>, Eleftheria Zeggini<sup>59</sup>, George V. Dedoussis<sup>35</sup>, Yik-Ying Teo<sup>33,80,81</sup>, Seang-Mei Saw<sup>33,82</sup>, Mads Melbye<sup>10,83,84</sup>, Harry Campbell<sup>32</sup>, James F. Wilson<sup>32,85</sup>, Martine Vrijheid<sup>29,30,31</sup>, Eco J. C. N. de Geus<sup>28,86</sup>, Dorret I. Boomsma<sup>28</sup>, Haja N. Kadarmideen<sup>87</sup>, Jens-Christian Holm<sup>12,61</sup>, Torben Hansen<sup>12</sup>, Sylvain Sebert<sup>57,88,89</sup>, Andrew T. Hattersley<sup>3</sup>, Lawrence J. Beilin<sup>90</sup>, John P. Newnham<sup>6</sup>, Craig E. Pennell<sup>6</sup>, Joachim Heinrich<sup>26,91</sup>, Linda S. Adair<sup>92</sup>, Judith B. Borja<sup>93,94</sup>, Karen L. Mohlke<sup>25</sup>, Johan G. Eriksson<sup>95,96,97</sup>, Elisabeth Widén<sup>22</sup>, Mika Kähönen<sup>98,99</sup>, Jorma S. Viikari<sup>100,101</sup>, Terho Lehtimäki<sup>20,21</sup>, Peter Vollenweider<sup>44</sup>, Klaus Bønnelykke<sup>16</sup>, Hans Bisgaard<sup>16</sup>, Dennis O. Mook-Kanamori<sup>15,102,103</sup>, Albert Hofman<sup>7,8</sup>, Fernando Rivadeneira<sup>7,8,67</sup>, André G. Uitterlinden<sup>7,8,67</sup>, Charlotta Pisinger<sup>104</sup>, Oluf Pedersen<sup>12</sup>, Christine Power<sup>105</sup>, Elina Hyppönen<sup>105,106,107</sup>, Nicholas J. Wareham<sup>4</sup>, Hakon Hakonarson<sup>13,23,108</sup>, Eleanor Davies<sup>41</sup>, Brian R. Walker<sup>68</sup>, Vincent W. V. Jaddoe<sup>7,8,9</sup>, Marjo-Riitta Järvelin<sup>88,89,109,110</sup>, Struan F. A. Grant<sup>13,23,108,111</sup>, Allan A. Vaag<sup>83,112,113</sup>, Debbie A. Lawlor<sup>55,56</sup>, Timothy M. Frayling<sup>3</sup>, George Davey Smith<sup>55,56</sup>, Andrew P. Morris<sup>1,114,115</sup>, Ken K. Ong<sup>4,116</sup>, Janine F. Felix<sup>7,8,9</sup>, Nicholas J. Timpson<sup>55,56</sup>, John R. B. Perry<sup>4</sup>, David M. Evans<sup>5,55,56</sup>, Mark I. McCarthy<sup>1,2,117</sup> & Rachel M. Freathy<sup>3,55</sup>

**Birth weight (BW) has been shown to be influenced by both fetal and maternal factors and in observational studies is reproducibly associated with future risk of adult metabolic diseases including type 2 diabetes (T2D) and cardiovascular disease<sup>1</sup>. These life-course associations have often been attributed to the impact of an adverse early life environment. Here, we performed a multi-ancestry genome-wide association study (GWAS) meta-analysis of BW in 153,781 individuals, identifying 60 loci where fetal genotype was associated with BW ( $P < 5 \times 10^{-8}$ ). Overall, approximately 15% of variance in BW was captured by assays of fetal genetic variation. Using genetic association alone, we found strong inverse genetic correlations between BW and systolic blood pressure ( $R_g = -0.22$ ,  $P = 5.5 \times 10^{-13}$ ), T2D ( $R_g = -0.27$ ,  $P = 1.1 \times 10^{-6}$ ) and coronary artery disease ( $R_g = -0.30$ ,  $P = 6.5 \times 10^{-9}$ ). In addition, using large-cohort datasets, we demonstrated that genetic factors were the major contributor to the negative covariance between BW and future cardiometabolic risk. Pathway analyses indicated that the protein products of genes within BW-associated regions were enriched for diverse processes including insulin signalling, glucose homeostasis, glycogen biosynthesis and chromatin remodelling. There was also enrichment of associations with BW in known imprinted regions ( $P = 1.9 \times 10^{-4}$ ). We demonstrate that life-course associations**

**between early growth phenotypes and adult cardiometabolic disease are in part the result of shared genetic effects and identify some of the pathways through which these causal genetic effects are mediated.**

We combined GWAS data for BW from 153,781 individuals representing multiple ancestries from 37 studies across three components (Extended Data Fig. 1 and Supplementary Table 1): (i) 75,891 individuals of European ancestry from 30 studies; (ii) 67,786 individuals of European ancestry from the UK Biobank; and (iii) 10,104 individuals of diverse ancestries (African American, Chinese, Filipino, Surinamese, Turkish and Moroccan) from six studies. Within each study, BW was Z-score transformed separately in males and females after excluding non-singletons and premature births and adjusting for gestational age where available. Genotypes were imputed using reference panels from the 1000 Genomes (1000G) Project<sup>2</sup> or combined 1000G and UK10K projects<sup>3</sup> (Supplementary Table 2). We performed quality control assessments to confirm that the distribution of BW was consistent across studies, irrespective of the data collection protocol, and confirmed that self-reported BW in the UK Biobank showed genetic and phenotypic associations consistent with those seen for measured BW in other studies<sup>4</sup> (Methods).

We identified 60 loci (of which 59 were autosomal) associated with BW at genome-wide significance ( $P < 5 \times 10^{-8}$ ) in either the European

ancestry or trans-ancestry meta-analyses (Extended Data Fig. 2a, Extended Data Table 1a and Supplementary Data; Methods). For lead single nucleotide polymorphisms (SNPs), we observed no heterogeneity in allelic effects between the three study components (Cochran's  $Q$  statistic  $P > 0.00083$ ) (Supplementary Table 3). We found that 53 of these loci were novel in that the lead SNP mapped  $>2$  Mb away from, and was independent ( $R^2 < 0.05$  in the European (EUR) component of 1000G) of, the seven previously reported BW signals<sup>5</sup>, all of which were confirmed in this larger analysis (Supplementary Table 4). Approximate conditional analysis in the European ancestry data indicated that three of these novel loci (near *ZBTB7B*, *HMGAI* and *PTCH1*) harboured multiple distinct association signals that attained genome-wide significance ( $P < 5 \times 10^{-8}$ ) (Methods, Supplementary Table 5 and Extended Data Fig. 3).

The lead variants for most signals mapped to non-coding sequences, and at only two loci, *ADRB1* (rs7076938;  $R^2 = 0.99$  with *ADRB1* G389R) and *NR1P1* (rs2229742, R448G), did the association data point to potential causal non-synonymous coding variants (Supplementary Table 6 and Methods). Lead SNPs for all but two loci (those mapping near *YKT6-GCK* and *SUZ12P1-CRLF3*) were common (minor allele frequency (MAF)  $\geq 5\%$ ) with individually modest effects on BW ( $\beta = 0.020$ – $0.053$  standard deviations (s.d.) per allele, equivalent to 10–26 g). This was despite the much-improved coverage of low-frequency variants in this study (compared to previous HapMap 2 imputed meta-analyses, ref. 5) reflecting imputation from larger, and more complete, reference panels (Extended Data Table 1b). Indeed, all but five of the common variant association signals were tagged by variants (EUR  $R^2 > 0.6$ ) in the HapMap 2 reference panel (Supplementary Tables 4, 5), indicating that most of the novel discoveries in the present study were driven by increased sample size<sup>5</sup>. Fine-mapping analysis yielded 14 regions in which fewer than ten variants contributed to the locus-specific credible sets that accounted for  $>99\%$  of the posterior probability of association (Methods and Supplementary Table 7). The greatest refinement was at *YKT6-GCK*, where the credible set included only the low frequency variant rs138715366, which maps intronic to *YKT6*. These credible-set variants collectively showed enrichment for overlap with DNaseI hypersensitivity sites, particularly those generated, by ENCODE, from fetal (4.2-fold, 95% CI 1.8–10.7) and neonatal tissues (4.9-fold, 1.8–11.0) (Supplementary Fig. 1, Supplementary Table 8 and Methods).

In combination, the 62 distinct genome-wide significant signals at the 59 autosomal loci explained at least  $2.0 \pm 1.1\%$  (standard error (s.e.)) of variance in BW (Supplementary Table 9 and Methods), which is similar in magnitude to that attributable to sex or maternal body mass index (BMI)<sup>5</sup>. However, the variance in BW captured collectively by all autosomal genotyped variants on the array was considerably larger, estimated at  $15.1 \pm 0.9\%$  in the UK Biobank (Methods). These figures are consistent with a large number of genetic variants with smaller effects contributing to variation in BW.

Associations between fetal genotype and BW could result from indirect effects of the maternal genotype influencing BW via the intrauterine environment, given the correlation ( $R \approx 0.5$ ) between maternal and fetal genotype. However, two lines of evidence indicated that variation in the fetal genome was the predominant driver of BW associations. First, an analysis of the global contribution of maternal versus fetal genetic variation, using a maternal genome-wide complex trait analysis (GCTA) model (ref. 6) (Methods) applied to 4,382 mother–child pairs, estimated that the child's genotype ( $\sigma_C^2 = 0.24 \pm 0.11$ ) made a larger contribution to BW variance than either the mother's genotype ( $\sigma_M^2 = 0.04 \pm 0.10$ ), or the covariance between the two ( $\sigma_{CM} = 0.04 \pm 0.08$ ). Second, when we compared the point estimates of the BW-effect size dependent on maternal genotype at each of the 60 loci (as measured in up to 68,254 women<sup>7</sup>) with those dependent on fetal genotype (using European ancestry data from 143,677 individuals in the present study), fetal variation had a greater impact than maternal variation at 93% of the loci (55 out of 60;

binomial  $P = 10^{-11}$ ) (Supplementary Table 10, Extended Data Figs 4, 5 and Methods). The power to further disentangle maternal and fetal contributions using analyses of fetal genotype which were conditional on maternal genotype was constrained by the limited sample size available ( $n = 12,909$  mother–child pairs) (Supplementary Table 11).

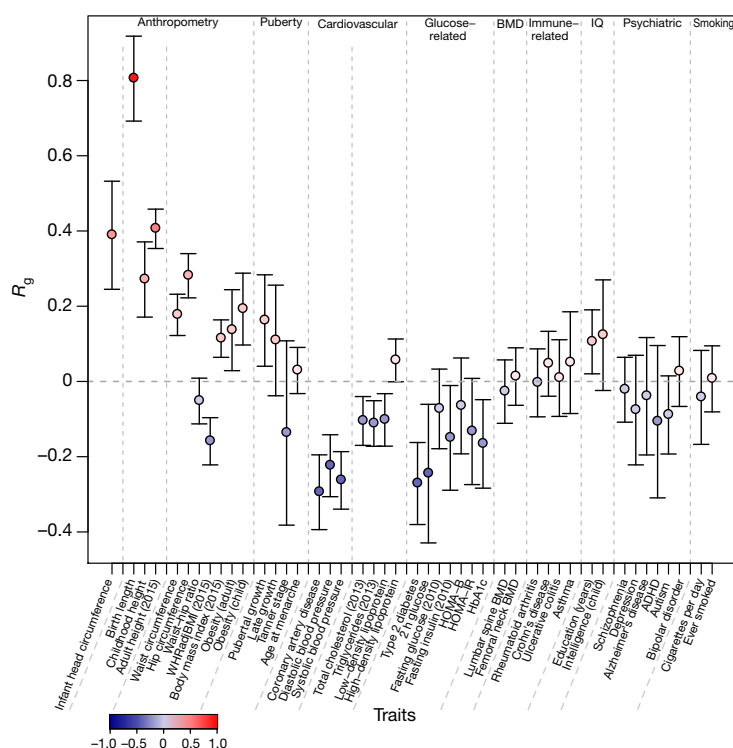
Collectively, these analyses provide evidence that the fetal genotype has a substantial impact on early growth, as measured by BW. We used these genetic associations to understand the causal relationships underlying observed associations between BW and disease, and to characterize the processes responsible.

To quantify the shared genetic contribution to BW and other health-related traits, we estimated their genetic correlations using linkage-disequilibrium score regression<sup>8</sup> (Methods). BW (in European ancestry samples) showed strong positive genetic correlations with anthropometric and obesity-related traits including birth length ( $R_g = 0.81$ ,  $P = 2.0 \times 10^{-44}$ ) and, in adults, height ( $R_g = 0.41$ ,  $P = 4.8 \times 10^{-52}$ ), waist circumference ( $R_g = 0.18$ ,  $P = 3.9 \times 10^{-10}$ ) and BMI ( $R_g = 0.11$ ,  $P = 7.3 \times 10^{-6}$ ). By contrast, BW showed inverse genetic correlations with indicators of adverse metabolic and cardiovascular health including coronary artery disease (CAD,  $R_g = -0.30$ ,  $P = 6.5 \times 10^{-9}$ ), systolic blood pressure (SBP,  $R_g = -0.22$ ,  $P = 5.5 \times 10^{-13}$ ) and T2D ( $R_g = -0.27$ ,  $P = 1.1 \times 10^{-6}$ ) (Fig. 1, Supplementary Table 12). The correlations between BW and adult cardiometabolic phenotypes are of similar magnitude, although directionally opposite, to the reported genetic correlations between adult BMI and those same cardiometabolic outcomes<sup>8</sup>. These findings support observational associations between a history of paternal T2D and lower BW (ref. 4), and establish more generally that the observed life-course associations between early growth and adult disease, at least in part, reflect the impact of shared genetic variants that influence both sets of phenotypes.

In an effort to estimate the extent of genetic contribution to these life-course associations, we first focused on data from the UK Biobank ( $n = 57,715$ ). For many of the traits for which data were available, genetic variation contributed substantially to the life-course relationship between BW and adult phenotypes, and in some cases appeared to be the major source of covariance between the traits. For example, we estimated that 85% (95% CI = 70–99%) of the negative covariance between BW and SBP was explained by shared genetic associations captured by directly genotyped SNPs (Supplementary Table 13, Methods and Supplementary Fig. 2). For continuous cardiometabolic measures, including lipids and fasting glycaemia, for which measures are not currently available in the UK Biobank, we used data from the Northern Finland Birth Cohort ( $n = 5,009$ ), and obtained similar results (Supplementary Table 13). However, these estimates were limited, not only by wide confidence intervals, but also by the assumption of a linear relationship between BW and each of the phenotypes and by the inability to explicitly model maternal genotypic effects. In other words, the inverse genetic correlations between BW and cardiometabolic traits may not exclusively reflect genetic effects mediated directly through the offspring, but also effects mediated by maternal genotype acting indirectly on the fetus via perturbation of the *in utero* environment. Nevertheless, these estimates indicate that a substantial proportion of the variance in cardiometabolic risk that correlates with BW can be attributed to the effects of common genetic variation.

To elucidate the biological pathways and processes underlying regulation of fetal growth, we first performed gene set enrichment analysis of our BW GWAS analysis using MAGENTA (Meta-Analysis Gene-set Enrichment of variaNT Associations, ref. 9) approach (Methods). Twelve pathways reached study-wide significance (false discovery rate, FDR  $< 0.05$ ), including pathways involved in metabolism (insulin signalling, glycogen biosynthesis and cholesterol biosynthesis), growth (IGF signalling and growth hormone pathway) and development (chromatin remodelling) (Extended Data Table 2a). Similar pathways were detected in a complementary analysis in which we analysed empirical protein–protein interaction (PPI) data identifying





**Figure 1 | Genome-wide genetic correlation between BW and a range of traits and diseases in later life.** Genetic correlation ( $R_g$ ) and corresponding s.e. (error bars) between BW and the traits displayed on the x axis were estimated using linkage-disequilibrium score regression (ref. 8). The genetic correlation estimates ( $R_g$ ) are colour coded according to their intensity and direction (red for positive and blue for inverse correlation). WHRadjBMI, waist-hip ratio adjusted for body mass index; HOMA-B/IR, homeostasis model assessment of beta-cell function/insulin resistance; HbA1c, haemoglobin A1c; BMD, bone mineral density; ADHD, attention deficit hyperactivity disorder. See Supplementary Table 12 for references for each of the traits and diseases displayed.

13 PPI network modules with marked ( $Z$  score  $> 5$ ) enrichment for BW-association scores (Extended Data Table 2b, Extended Data Fig. 6a, b and Methods). The proteins within these modules were themselves enriched for diverse processes related to metabolism, growth and development (Extended Data Fig. 6a, b).

We also observed enrichment of BW association signals across the set of 77 imprinted genes defined by the Genotype-Tissue Expression (GTEx) project (ref. 10) ( $P = 1.9 \times 10^{-4}$ ; Extended Data Table 2a and Supplementary Table 14). Such enrichment is consistent with the 'parental conflict' hypothesis regarding the allocation of maternal resources to the fetus<sup>11</sup>. Although the role of imprinted genes in fetal growth has been described in animal models and rare human disorders<sup>12</sup>, these data provide a large-scale, systematic indication of their contribution to normal variation in BW. Of the 60 genome-wide significant loci, two (*INS-IGF2* and *RB1*) fall within (or near) imprinted regions (Extended Data Fig. 2b), with a noteworthy third signal at *DLK1* (previously fetal antigen-1;  $P = 5.6 \times 10^{-8}$ ). Parent-of-origin specific analyses to further investigate these individual loci (comparing heterozygote versus homozygote BW variance in 57,715 unrelated individuals, and testing BW associations with paternal versus maternal alleles in 4,908 mother-child pairs; see Methods) proved, despite these sample sizes, to be underpowered (Extended Data Fig. 7 and Supplementary Tables 15, 16).

Many of the genome-wide signals for BW detected here are also established genome-wide association signals for a wide variety of cardiometabolic traits (Fig. 2). These include the BW signals near *CDKAL1*, *ADCY5*, *HHEX-IDE* and *ANK1* (also genome-wide significant for T2D), *NT5C2* (for blood pressure, CAD and BMI) and *ADRB1* (for blood pressure). We used two approaches to understand whether this pattern of adult trait association represented a generic property of BW-associated loci or reflected heterogeneous mechanisms linking BW to adult disease.

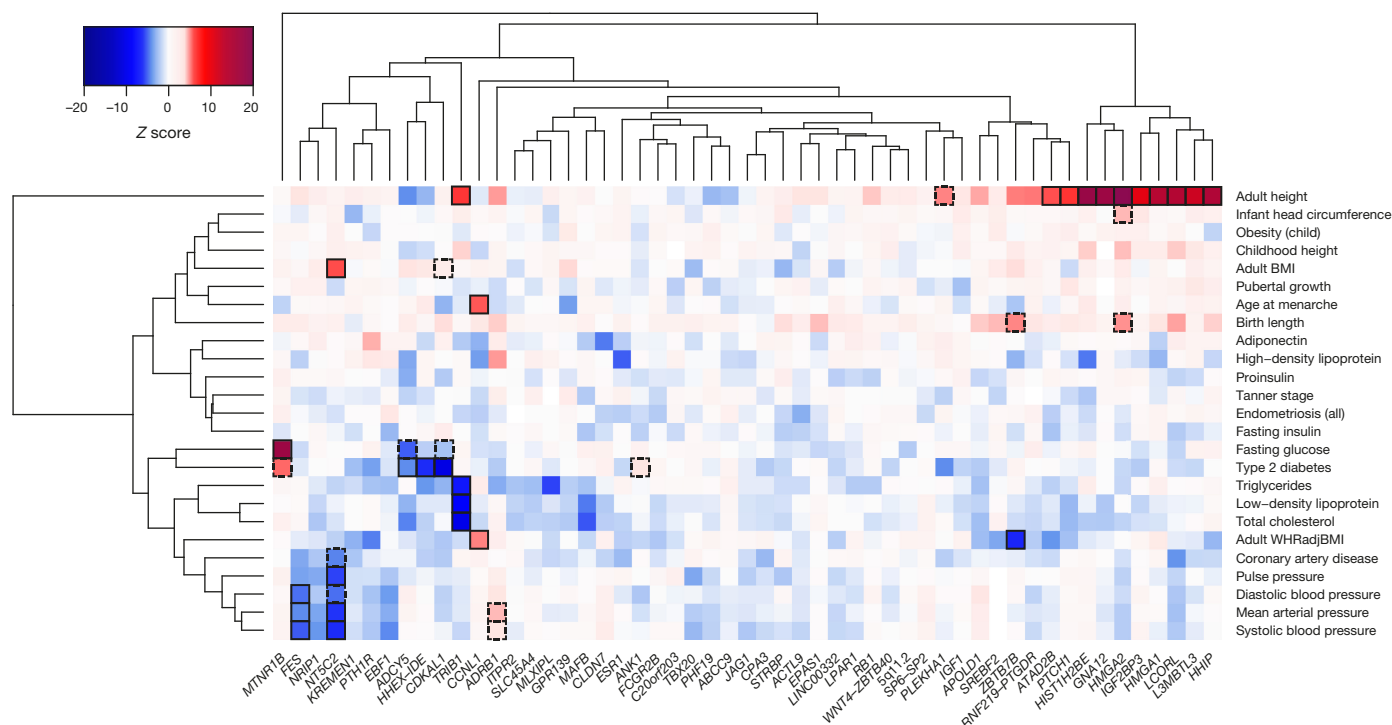
First, we applied unsupervised hierarchical clustering (Methods) to the non-BW trait association statistics for the 60 significant BW loci. The resultant heat map showed the heterogeneity of locus-specific effect sizes across the range of adult traits (Fig. 2 and Supplementary Table 17). For example, it revealed that the associations between BW-raising alleles and increased adult height are concentrated amongst a subset

of loci including *HHIP* and *GNA12*, and highlighted particularly strong associations with lipid traits for variants at the *TRIB1* and *MAFB* loci.

Second, we constructed trait-specific 'point-of-contact' (PoC) PPI networks from proteins represented in both the global BW PPI network and equivalent PPI networks generated for each of the adult traits (Methods and Extended Data Figs 6c–e). We reasoned that these PoC PPI networks would be enriched for the specific proteins mediating the observed links between BW and adult traits, generating hypotheses that are amenable to subsequent empirical validation. To highlight processes implicated in specific BW-trait associations, we overlaid these PoC PPI with the top 50 pathways that were over-represented in the global BW PPI network. These analyses revealed, for example, that proteins in the Wnt canonical signalling pathway were detected in the PoC PPI network only for blood pressure traits. We used these PPI overlaps to highlight the specific transcripts within BW GWAS loci that were likely to mediate the mechanistic links. For example, the overlap between the Wnt signalling pathway and the PoC PPI network for the intersection of BW and blood pressure-related traits implicated *FZD9* as the likely effector gene at the *MLXIPL* BW locus (Extended Data Fig. 6d and Supplementary Table 6).

We focused our more detailed investigation of the mechanistic links between early growth and adult traits on two phenotypic areas: arterial blood pressure and T2D/glycaemia. Across both the overall GWAS and specifically among the 60 significant BW loci, most BW-raising alleles were associated with reduced blood pressure (Figs 1, 2); the strongest inverse associations were seen for the loci near *NT5C2*, *FES*, *NRI1*, *EBF1* and *PTH1R*. However, we also observed locus-specific heterogeneity in the genetic relationships between blood pressure and BW: the SBP-raising allele at *ADRB1*<sup>13</sup> is associated with higher, rather than lower, BW (Extended Data Fig. 8a). When we considered the reciprocal relationship, that is, the effects on BW of blood-pressure-raising alleles at 30 reported loci for SBP<sup>13,14</sup>, there was an excess of associations (5 out of 30 with lower BW at  $P < 0.05$ ; binomial  $P = 0.0026$ ; Extended Data Fig. 8a). To dissect maternal and fetal genotype effects at these loci, we tested the impact on BW of a risk score generated from the 30 SBP SNPs, restricted to the untransmitted maternal haplotype score<sup>15</sup> in a set of 5,201 mother-child pairs. Analysis of these loci indicated that maternal genotype effects on the intrauterine environment probably





**Figure 2 | Hierarchical clustering of BW loci based on similarity of overlap with adult diseases, metabolic and anthropometric traits.** For the lead SNP at each BW locus (*x* axis), Z scores (aligned to BW-raising allele) were obtained from publicly available GWAS for various traits (*y* axis; see Supplementary Table 17). A positive Z score (red) indicates a positive association between the BW-raising allele and the outcome trait,

while a negative Z score (blue) indicates an inverse association. BW loci and traits were clustered according to the Euclidean distance amongst Z scores (see Methods). Squares are outlined with a solid black line if the BW locus is significantly ( $P < 5 \times 10^{-8}$ ) associated with the trait in publicly available GWAS, or with a dashed line if reported significant elsewhere.

contribute to the inverse genetic correlation between SBP and BW (Methods and Supplementary Table 18), and was consistent with the results of a wider study of >30,000 women which demonstrated associations between a maternal genetic score for SBP (conditional on fetal genotype) and lower offspring BW<sup>16</sup>.

The blood-pressure-raising allele with the largest BW-lowering effect mapped to the *NT5C2* locus (index variant for BW, rs74233809,  $R^2 = 0.98$  with index variant for blood pressure, rs11191548; ref. 14) and was also associated with lower adult BMI ( $R^2 = 0.99$  with rs11191560; ref. 17). The BW-lowering allele at rs74233809 is a proxy for a recently described<sup>18</sup> functional variant in the nearby *CYP17A1* gene ( $R^2 = 0.92$  with rs138009835). The *CYP17A1* gene encodes the cytochrome P450c17 $\alpha$  enzyme CYP17 (ref. 19), which catalyses key steps in steroidogenesis that determine the balance between mineralocorticoid, glucocorticoid and androgen synthesis. This variant has been shown to alter transcriptional efficiency *in vitro* and is associated with increased urinary tetrahydroaldosterone excretion<sup>18</sup>. *CYP17A1* is expressed in fetal adrenal glands and testes from early gestation<sup>20</sup> as well as in the placenta<sup>21</sup>. These data suggest that variation in *CYP17A1* expression contributes to the observational association between low BW and adult hypertension<sup>22</sup>.

When we analysed 45 loci associated with CAD<sup>23</sup>, the inverse genetic correlation between CAD and BW was concentrated amongst the five CAD loci with primary blood pressure associations. This suggests that genetic determinants of blood pressure play a leading role in mediating the life-course associations between BW and CAD (Extended Data Fig. 8b, e).

Linkage-disequilibrium score regression analyses demonstrated overall inverse genetic correlation between lower BW and elevated risk of T2D (Fig. 1). However, the locus-specific heat map indicates a heterogeneous pattern across individual loci (Fig. 2). To explore this further, we tested the 84 reported T2D loci<sup>24</sup> for association with BW. Some T2D risk alleles (such as those at *ADCY5*, *CDKAL1* and *HHEX-IDE*) were strongly associated with lower BW, while others (including *ANK1* and

*MTNR1B*) were associated with higher BW (Extended Data Fig. 8c). This was in contrast with the BW effects of 422 known height loci<sup>25</sup> (Extended Data Fig. 8d), which showed a strong positive correlation consistent with the overall genetic correlation between height and BW, indicating that the growth effects of many height loci start prenatally and persist into adulthood.

The contrasting associations of T2D-risk alleles with both higher and lower BW probably reflect the differential impacts, across loci, of variation in the maternal and fetal genomes. Observational data link paternal diabetes with lower offspring BW<sup>4</sup>, indicating that the inheritance of T2D risk alleles by the fetus tends, in line with the linkage-disequilibrium score regression analysis, to reduce growth. These relationships are consistent with the precepts of the 'fetal insulin hypothesis'<sup>26</sup> and reflect the potential for reduced insulin secretion and/or signalling to lead to both reduced fetal growth and, many decades later, enhanced predisposition to T2D. In line with this, the inferred paternal transmitted haplotype score generated from the 84 T2D risk variants was associated with lower BW ( $P = 0.045$ ) in 5,201 mother-child pairs (Methods and Supplementary Table 18). In contrast, maternal diabetes is observationally associated with higher offspring BW<sup>4</sup>, reflecting the ability of maternal hyperglycaemia to stimulate fetal insulin secretion. The contribution of genotype-dependent maternal hyperglycaemia to BW is in line with the evidence, from a recent study, that maternal genotype scores for fasting glucose and T2D (conditional on fetal genotype) were causally associated with higher offspring BW<sup>16</sup>. It is also consistent with the observation that a subset of glucose-raising alleles is associated with higher BW<sup>7</sup>. For example, the T2D-risk variant at *MTNR1B* (which also has a marked effect on fasting glucose levels in non-diabetic individuals<sup>27,28</sup>) was amongst the subset of BW loci (5 out of 60) for which the BW effect attributable to maternal genotype exceeded that associated with the fetal genotype (maternal:  $\beta = 0.048$ ,  $P = 5.1 \times 10^{-15}$ ; fetal:  $\beta = 0.023$ ,  $P = 2.9 \times 10^{-8}$ ) (Supplementary Table 10 and Extended Data Figs 4, 5). Thus, both maternal and fetal genetic effects connect BW to later T2D risk, albeit acting in opposing

directions. When we categorized T2D loci using a classification of physiological functions derived from their effects on related glycaemic and anthropometric traits<sup>27</sup>, we found that T2D-risk alleles associated with lower BW were those typically characterized by reduced insulin processing and secretion without detectable changes in fasting glucose (the 'Beta Cell' cluster in Extended Data Fig. 8f).

The *YTK6* signal at rs138715366 is notable not only because the genetic data indicate that a single low-frequency non-coding variant is driving the association signal (see above) but also because of the proximity of this signal to *GCK*. Rare coding variants in glucokinase are causal for a form of monogenic hyperglycaemia and lead to large reductions in BW when parental alleles are passed on to their offspring<sup>29</sup>. In addition, common non-coding variants nearby are implicated in T2D risk and fasting hyperglycaemia<sup>28</sup>. However, the latter variants are conditionally independent of rs138715366 (Supplementary Table 19) and show no comparable association with lower BW. Either rs138715366 acts through effector transcripts other than *GCK*, or the impact of the low-frequency SNP near *YTK6* on *GCK* expression involves tissue- and/or temporal-specific variation in regulatory impact.

In conclusion, we have identified 60 genetic loci associated with BW and used them to gain insights into the aetiology of fetal growth and into well-established, but until now poorly understood, life-course disease associations. The evidence that the relationship between early growth and later metabolic disease has an appreciable genetic component contrasts with, but is not necessarily incompatible with, the emphasis on adverse early environmental events highlighted by the fetal origins hypothesis<sup>1</sup>. As we have shown, these genetic effects reflect variation in both the fetal and the maternal genome: the impact of the latter on the offspring's predisposition to adult disease could be mediated, at least in part, through perturbation of the antenatal and early life environment. Future mechanistic and genetic studies should support reconciliation between these alternative, but complementary, explanations for the far-reaching life-course associations that exist between events in early life and predisposition to cardiometabolic disease several decades later.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 4 February; accepted 2 September 2016.**

**Published online 28 September 2016.**

- Barker, D. J. The developmental origins of chronic adult disease. *Acta Paediatr. Suppl.* **93**, 26–33 (2004).
- The 1000 Genomes Project Consortium An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- The UK10K Project Consortium The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
- Tyrrrell, J. S., Yaghootkar, H., Freathy, R. M., Hattersley, A. T. & Frayling, T. M. Parental diabetes and birthweight in 236,030 individuals in the UK Biobank study. *Int. J. Epidemiol.* **42**, 1714–1723 (2013).
- Horikoshi, M. *et al.* New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat. Genet.* **45**, 76–82 (2013).
- Eaves, L. J., Pourcain, B. S., Smith, G. D., York, T. P. & Evans, D. M. Resolving the effects of maternal and offspring genotype on dyadic outcomes in genome wide complex trait analysis ("M-GCTA"). *Behav. Genet.* **44**, 445–455 (2014).
- Feenstra, B. *et al.* Maternal genome-wide association study identifies a fasting glucose variant associated with offspring birth weight. Preprint at: <http://biorxiv.org/content/early/2015/12/11/034207> (2015).
- Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
- Segrè, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).
- Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–936 (2015).
- Haig, D. & Westoby, M. Parent-specific gene expression and the triploid endosperm. *Am. Nat.* **134**, 147–155 (1989).
- Peters, J. The role of genomic imprinting in biology and disease: an expanding view. *Nat. Rev. Genet.* **15**, 517–530 (2014).

- Johnson, T. *et al.* Blood pressure loci identified with a gene-centric array. *Am. J. Hum. Genet.* **89**, 688–700 (2011).
- International Consortium for Blood Pressure Genome-Wide Association Studies *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).
- Zhang, G. *et al.* Assessing the causal relationship of maternal height on birth size and gestational age at birth: a Mendelian randomization analysis. *PLoS Med.* **12**, e1001865 (2015).
- Tyrrrell, J. *et al.* Genetic evidence for causal relationships between maternal obesity-related traits and birth weight. *J. Am. Med. Assoc.* **315**, 1129–1140 (2016).
- Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
- Diver, L. A. *et al.* Common polymorphisms at the *CYP17A1* locus associate with steroid phenotype: support for blood pressure genome-wide association study signals at this locus. *Hypertension* **67**, 724–732 (2016).
- Picado-Leonard, J. & Miller, W. L. Cloning and sequence of the human gene for P450c17 (steroid 17 alpha-hydroxylase/17,20 lyase): similarity with the gene for P450c21. *DNA* **6**, 439–448 (1987).
- Pezzi, V., Mathis, J. M., Rainey, W. E. & Carr, B. R. Profiling transcript levels for steroidogenic enzymes in fetal tissues. *J. Steroid Biochem. Mol. Biol.* **87**, 181–189 (2003).
- Escobar, J. C., Patel, S. S., Beshay, V. E., Suzuki, T. & Carr, B. R. The human placenta expresses CYP17 and generates androgens *de novo*. *J. Clin. Endocrinol. Metab.* **96**, 1385–1392 (2011).
- Reynolds, R. M. *et al.* Programming of hypertension: associations of plasma aldosterone in adult men and women with birthweight, cortisol, and blood pressure. *Hypertension* **53**, 932–936 (2009).
- CARDIoGRAMplusC4D Consortium *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat. Genet.* **45**, 25–33 (2013).
- DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
- Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
- Hattersley, A. T. & Tooke, J. E. The fetal insulin hypothesis: an alternative explanation of the association of low birth weight with diabetes and vascular disease. *Lancet* **353**, 1789–1792 (1999).
- Dimas, A. S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63**, 2158–2171 (2014).
- Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
- Hattersley, A. T. *et al.* Mutations in the glucokinase gene of the fetus result in reduced birth weight. *Nat. Genet.* **19**, 268–270 (1998).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** Full acknowledgements and supporting grant details can be found in the Supplementary Information.

**Author Contributions** Core analyses and writing: M.H., R.N.B., F.R.D., N.M.W., M.N.K., J.F.-T., N.R.v.Z., K.J.G., A.P.M., K.K.O., J.F.F., N.J.T., J.R.P., D.M.E., M.I.M., R.M.F. Statistical analysis in individual studies: M.H., R.N.B., F.R.D., N.M.W., M.N.K., B.F., N.G., J.P.B., D.P.S., R.L.-G., T.S.A., E.K., R.R., L.-P.L., D.L.C., Y.W., E.T., C.A.W., C.T.H., J.-J.H., N.V.-T., P.K.J., E.T.H.B., I.N., N.P., A.M., E.M.v.L., R.J., V.L.a., M.N., J.M.M., S.E.J., P.-R.L., K.S.R., M.A.T., J.T., A.R.W., H.Y., D.M.S., I.P., K.Pan., X.W., L.C., F.G., K.E.S., M.Mu., E.V.R.A., Z.K., S.B.-G., F.S., D.T., J.W., C.M.-G., N.R.R., E.Z., G.V.D., Y.-Y.T., H.N.K., A.P.M., J.F.F., N.J.T., J.R.P., D.M.E., R.M.F. GWAS look-up in unpublished datasets: K.T.Z., N.R., D.R.N., R.C.W.M., C.H.T.T., W.H.T., S.K.G., F.J.v.R. Sample collection and data generation in individual studies: F.R.D., M.N.K., B.F., N.G., J.P.B., D.P.S., R.L.-G., R.R., L.-P.L., J.-J.H., I.N., E.M.v.L., M.B., P.M.-V., A.J.B., L.P., P.K., M.A., S.M.W., F.G., C.E.v.B., G.W., E.V.R.A., C.E.F., C.T., C.M.T., M.Sta., Z.K., D.M.H., M.V.H., H.G.d.H., F.R.R., C.M.-G., S.M.R., G.H., G.M., N.R.R., C.J.G., C.L., J.L., R.A.S., J.H.Z., F.D.M., W.L.L.Jr., A.T., M.Stu., V.Li., T.A.L., C.M.v.D., A.K., T.I.S., H.N., K.Pah., O.T.R., E.Z., G.V.D., S.-M.S., M.Me., H.C., J.F.W., M.V., J.-C.H., T.H., S.S., L.J.B., J.P.N., C.E.P., L.S.A., J.B.B., K.L.M., J.G.E., E.E.W., M.K., J.S.V., T.L., P.V., K.B., H.B., D.O.M.-K., A.H., F.R., A.G.U., C.Pi., O.P., C.Po., E.H., N.J.W., H.H., V.W.J., M.-R.J., S.F.G., A.A.V., T.M.F., A.P.M., K.K.O., N.J.T., J.R.P., M.I.M., R.M.F.

**Author Information** Summary statistics from the meta-analyses are available at <http://egg-consortium.org/>. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.I.M. (mark.mccarthy@drli.ox.ac.uk) or R.M.F. (r.freathy@ex.ac.uk).

**Reviewer Information** *Nature* thanks J. Whitfield and the other anonymous reviewer(s) for their contribution to the peer review of this work.

- <sup>1</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
- <sup>2</sup>Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LE, UK. <sup>3</sup>Institute of Biomedical and Clinical Science, University of Exeter Medical School, University of Exeter, Royal Devon and Exeter Hospital, Exeter EX2 5DW, UK. <sup>4</sup>MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge CB2 0QQ, UK. <sup>5</sup>The University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland 4102, Australia. <sup>6</sup>School of Women's and Infants' Health, The University of Western Australia, Perth, Western Australia 6009, Australia. <sup>7</sup>The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands.
- <sup>8</sup>Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands. <sup>9</sup>Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands. <sup>10</sup>Department of Epidemiology Research, Statens Serum Institut, Copenhagen DK-2300, Denmark. <sup>11</sup>Department of Pediatrics, University of California San Diego, La Jolla, San Diego, California 92093, USA. <sup>12</sup>The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen DK-2100, Denmark. <sup>13</sup>Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>14</sup>Population Health Research Institute, St George's University of London, Cranmer Terrace, London SW17 0RE, UK. <sup>15</sup>Department of Clinical Epidemiology, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands. <sup>16</sup>COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, 2820 Gentofte, Denmark. <sup>17</sup>Steno Diabetes Center, Gentofte DK-2820, Denmark. <sup>18</sup>Department of Computational Biology, University of Lausanne, Lausanne 1011, Switzerland. <sup>19</sup>Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland.
- <sup>20</sup>Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland. <sup>21</sup>Department of Clinical Chemistry, University of Tampere School of Medicine, Tampere 33014, Finland. <sup>22</sup>Institute for Molecular Medicine, Finland (FIMM), University of Helsinki, Helsinki FI-00100, Finland. <sup>23</sup>Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>24</sup>Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>25</sup>Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599, USA. <sup>26</sup>Institute of Epidemiology I, Helmholtz Zentrum München - German Research Center for Environmental Health, 85764 Neuherberg, Germany. <sup>27</sup>Division of Metabolic and Nutritional Medicine, Dr. von Hauner Children's Hospital, University of Munich Medical Center, 80337 Munich, Germany. <sup>28</sup>Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit, Amsterdam 1081 BT, the Netherlands. <sup>29</sup>ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona 08003, Spain. <sup>30</sup>Universitat Pompeu Fabra (UPF), Barcelona 08002, Spain. <sup>31</sup>CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain. <sup>32</sup>Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh EH8 9AG, UK. <sup>33</sup>Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore 119077, Singapore. <sup>34</sup>William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK. <sup>35</sup>Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens 17671, Greece. <sup>36</sup>Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20014, Finland. <sup>37</sup>Institute of Biomedicine, Physiology, University of Eastern Finland, Kuopio FI-70211, Finland. <sup>38</sup>KUL – University of Leuven, Department of Neurosciences, Leuven 3000, Belgium. <sup>39</sup>Translational Immunology Laboratory, VIB, Leuven 3000, Belgium. <sup>40</sup>Department of Preventive Medicine, Division of Biostatistics, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA. <sup>41</sup>Institute of Cardiovascular & Medical Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, UK. <sup>42</sup>Endometriosis CaRe Centre, Nuffield Department of Obstetrics & Gynaecology, University of Oxford, Oxford OX3 9DU, UK. <sup>43</sup>Center for Genomic Regulation (CRG), Barcelona 08003, Spain. <sup>44</sup>Department of Internal Medicine, Internal Medicine, Lausanne University Hospital (CHUV), Lausanne 1011, Switzerland. <sup>45</sup>Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona 08034, Spain. <sup>46</sup>Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland 4000, Australia. <sup>47</sup>Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. <sup>48</sup>Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. <sup>49</sup>Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. <sup>50</sup>Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China. <sup>51</sup>Department of Human Genetics and Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan 48109, USA. <sup>52</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA. <sup>53</sup>Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. <sup>54</sup>European Centre for Environment and Human Health, University of Exeter, Truro TR1 3HD, UK. <sup>55</sup>Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol BS8 2BN, UK. <sup>56</sup>School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK. <sup>57</sup>Department of Genomics of Common Disease, School of Public Health, Imperial College London, London SW7 2AZ, UK. <sup>58</sup>FB Adiposity Diseases, University of Leipzig, 04103 Leipzig, Germany. <sup>59</sup>Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK. <sup>60</sup>FISABIO–Universitat Jaume I–Universitat de València, Joint Research Unit of Epidemiology and Environmental Health, Valencia 46020, Spain. <sup>61</sup>The Children's Obesity Clinic, Department of Pediatrics, Copenhagen University Hospital Holbæk, Holbæk DK-4300, Denmark. <sup>62</sup>Institute of Social and Preventive Medicine, Lausanne University Hospital (CHUV), Lausanne 1010, Switzerland. <sup>63</sup>Danish Center for Neonatal Screening, Statens Serum Institute, Copenhagen DK-2300, Denmark. <sup>64</sup>Department for Congenital Disorders, Statens Serum Institute, Copenhagen DK-2300, Denmark. <sup>65</sup>Computer Sciences Department, Barcelona Supercomputing Center, Barcelona 08034, Spain. <sup>66</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain. <sup>67</sup>Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands. <sup>68</sup>BHF Centre for Cardiovascular Science, University of Edinburgh, Queen's Medical Research Institute, Edinburgh EH16 4JT, UK. <sup>69</sup>Department of Medicine, Division of Endocrinology, Metabolism, and Molecular Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA. <sup>70</sup>Medical Department, University of Leipzig, 04103 Leipzig, Germany. <sup>71</sup>Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, Kuopio FI-70029, Finland. <sup>72</sup>Kuopio Research Institute of Exercise Medicine, Kuopio FI-70100, Finland. <sup>73</sup>Pediatric Research Center, Department of Women's & Child Health, University of Leipzig, 04103 Leipzig, Germany. <sup>74</sup>Novo Nordisk Foundation Center for Basic Metabolic Research and Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark. <sup>75</sup>Institute of Preventive Medicine, Bispebjerg and Frederiksberg Hospital, The Capital Region, Copenhagen DK-2000, Denmark. <sup>76</sup>Department of Pediatrics, Turku University Hospital, Turku 20521, Finland. <sup>77</sup>Department of Physiology, University of Turku, Turku 20014, Finland. <sup>78</sup>Paavo Nurmi Centre, Sports and Exercise Medicine Unit, Department of Physical Activity and Health, Turku 20014, Finland. <sup>79</sup>Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku 20521, Finland. <sup>80</sup>Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore. <sup>81</sup>Life Sciences Institute, National University of Singapore, Singapore 117456, Singapore. <sup>82</sup>Singapore Eye Research Institute, Singapore 168751, Singapore. <sup>83</sup>Department of Clinical Medicine, University of Copenhagen, Copenhagen DK-2200, Denmark. <sup>84</sup>Department of Medicine, Stanford School of Medicine, Stanford, California 94305, USA. <sup>85</sup>MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. <sup>86</sup>EMGO Institute for Health and Care Research, VU University and VU University Medical Center, Amsterdam 1081 HV, the Netherlands. <sup>87</sup>Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg C DK-1870, Denmark. <sup>88</sup>Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu 90014, Finland. <sup>89</sup>Biocenter Oulu, University of Oulu, Oulu 90014, Finland. <sup>90</sup>School of Medicine and Pharmacology, Royal Perth Hospital Unit, The University of Western Australia, Perth, Western Australia 6000, Australia. <sup>91</sup>Institute and Outpatient Clinic for Occupational, Social and Environmental Medicine, Inner City Clinic, University Hospital Munich, Ludwig Maximilian University of Munich, 80336 Munich, Germany. <sup>92</sup>Department of Nutrition, University of North Carolina, Chapel Hill, North Carolina 27599, USA. <sup>93</sup>USC-Office of Population Studies Foundation, Inc., University of San Carlos, Cebu City 6000, Philippines. <sup>94</sup>Department of Nutrition and Dietetics, University of San Carlos, Cebu City 6000, Philippines. <sup>95</sup>National Institute for Health and Welfare, Helsinki 00271, Finland. <sup>96</sup>Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki 00014, Finland. <sup>97</sup>Folkhälsan Research Center, Helsinki 00250, Finland. <sup>98</sup>Department of Clinical Physiology, Tampere University Hospital, Tampere 33521, Finland. <sup>99</sup>Department of Clinical Physiology, University of Tampere School of Medicine, Tampere 33014, Finland. <sup>100</sup>Division of Medicine, Turku University Hospital, Turku 20521, Finland. <sup>101</sup>Department of Medicine, University of Turku, Turku 20014, Finland. <sup>102</sup>Department of Public Health and Primary Care, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands. <sup>103</sup>Epidemiology Section, BESC Department, King Faisal Specialist Hospital and Research Centre, Riyadh 12713, Saudi Arabia. <sup>104</sup>Research Center for Prevention and Health Capital Region, Center for Sundhed, Rigshospitalet – Glostrup, Copenhagen University, Glostrup DK-2600, Denmark. <sup>105</sup>Population, Policy and Practice, UCL Institute of Child Health, University College London, London WC1N 1EH, UK. <sup>106</sup>Centre for Population Health Research, School of Health Sciences, and Sansom Institute, University of South Australia, Adelaide, South Australia 5001, Australia. <sup>107</sup>South Australian Health and Medical Research Institute, Adelaide, South Australia 5000, Australia. <sup>108</sup>Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. <sup>109</sup>Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment & Health, School of Public Health, Imperial College London, London SW7 2AZ, UK. <sup>110</sup>Unit of Primary Care, Oulu University Hospital, Oulu 90220, Finland. <sup>111</sup>Division of Endocrinology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. <sup>112</sup>Department of Endocrinology, Rigshospitalet, Copenhagen DK-2100, Denmark. <sup>113</sup>AstraZeneca, Innovative Medicines and Early Development | Early Clinical Development, Mölndal 431 83, Sweden. <sup>114</sup>Department of Biostatistics, University of Liverpool, Liverpool L69 3GA, UK. <sup>115</sup>Estonian Genome Center, University of Tartu, Tartu 50090, Estonia. <sup>116</sup>Department of Paediatrics, University of Cambridge, Cambridge CB2 0QQ, UK. <sup>117</sup>Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Churchill Hospital, Oxford OX3 7LE, UK.
- \*These authors contributed equally to this work.  
†A list of consortium members appears in the Supplementary Information.  
‡Deceased.  
§These authors jointly supervised this work.



## METHODS

**Ethics statement.** All human research was approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. All participants provided written informed consent. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the local Research Ethics Committees.

**Study-level analyses.** No statistical methods were used to predetermine sample size: to maximise power to detect association signals, we set out to collect the largest possible set of samples for which the combination of genome-wide genotyping data and reliable measures of BW could be made available for analysis. Within each study, BW was collected from a variety of sources, including measurements at birth by medical practitioners, obstetric records, medical registers, interviews with the mother and self-report as adults (Supplementary Table 1). BW was Z-score transformed separately in males and females. Individuals with extreme BW ( $>5$  s.d. from the sex-specific study mean), monozygotic or polyzygotic siblings, or preterm births (gestational age  $<37$  weeks, where this information was available) were excluded from downstream association analyses (Supplementary Table 1).

Within each study, stringent quality control of the GWAS genotype scaffold was carried out before imputation (Supplementary Table 2). Each scaffold was then pre-phased and imputed<sup>30,31</sup> up to reference panels from the 1000G project<sup>2</sup> or the combined 1000G and UK10K projects<sup>3</sup> (Supplementary Table 2). Association of BW with each variant passing established GWAS quality control filters<sup>32</sup> was tested in a linear regression framework, under an additive model for the allelic effect, after adjustment for study-specific covariates, including gestational age, where available (Supplementary Table 2). Where necessary, population structure was accounted for by adjustment for axes of genetic variation from principal components analysis<sup>33</sup> and subsequent genomic control correction<sup>34</sup>, or inclusion of a genetic relationship matrix in a mixed model<sup>35</sup> (Supplementary Table 2). We calculated the genomic control inflation factor ( $\lambda$ ) in each study to confirm that study-level population structure was accounted for before meta-analysis.

**Preparation, quality control and genetic analysis in UK Biobank samples.** UK Biobank phenotype data were available for 502,655 participants<sup>36</sup>. All participants in the UK Biobank were asked to recall their BW, of which 279,971 did so at either the baseline or follow-up assessment visit. Of these, 7,686 participants reported being part of multiple births and were excluded from downstream analyses. Ancestry checks, based on self-reported ancestry, resulted in the exclusion of 8,998 additional participants reported not to be white European. Of those individuals reporting BW at baseline and follow-up assessments, 393 were excluded because the two reported values differed by more than 0.5 kg. For those reporting different values ( $\leq 0.5$  kg) between baseline and follow-up, we took the baseline measure forward for downstream analyses. We then excluded 36,716 individuals reporting values  $<2.5$  kg or  $>4.5$  kg as implausible for live term births before 1970. In total 226,178 participants had data relating to BW that matched these inclusion criteria.

Genotype data from the May 2015 release were available for a subset of 152,249 participants from UK Biobank. In addition to the quality control metrics performed centrally by UK Biobank, we defined a subset of 'white European' ancestry samples using a  $K$ -means ( $K=4$ ) clustering approach based on the first four genetically determined principal components. A maximum of 67,786 individuals (40,425 females and 27,361 males) with genotype and valid BW measures were available for downstream analyses. We tested for association with BW, assuming an additive allelic effect, in a linear mixed model implemented in BOLT-LMM (ref. 37) to account for cryptic population structure and relatedness. Genotyping array was included as a binary covariate in all models. Total chip heritability (that is, the variance explained by all autosomal polymorphic genotyped SNPs passing quality control) was calculated using restricted maximum likelihood (REML) implemented in BOLT-LMM (ref. 37). We additionally analysed the association between BW and directly genotyped SNPs on the X chromosome: for this analysis, we used 57,715 unrelated individuals with BW available and identified by UK Biobank as white British. We excluded SNPs with evidence of deviation from Hardy-Weinberg equilibrium ( $P < 1 \times 10^{-6}$ ), MAF  $< 0.01$  or overall missing rate  $> 0.015$ , resulting in 19,423 SNPs for analysis in Plink v1.07 (<http://pngu.mgh.harvard.edu/purcell/plink/>)<sup>38</sup>, with the first five ancestry principal components as covariates.

In both the full UK Biobank sample and our refined sample, we observed that BW was associated with sex, year of birth and maternal smoking ( $P < 0.0015$ , all in the expected directions), confirming more comprehensive previous validation of self-reported BW<sup>4</sup>. We additionally verified that BW associations with lead SNPs at seven established loci<sup>5</sup> based on self-report in UK Biobank were consistent with those previously published.

**European ancestry meta-analysis.** The European ancestry meta-analysis consisted of two components: (i) 75,891 individuals from 30 GWAS from Europe, USA and Australia; and (ii) 67,786 individuals of white European origin from the UK Biobank. In the first component, we combined sex-specific BW association

summary statistics across studies in a fixed-effects meta-analysis, implemented in GWAMA (ref. 39) and applied a second round of genomic control<sup>34</sup> ( $\lambda_{GC} = 1.001$ ). Subsequently, we combined association summary statistics from this component with the UK Biobank in a European ancestry fixed-effects meta-analysis, implemented in GWAMA (ref. 39). Variants failing GWAS quality control filters in the UK Biobank, reported in less than 50% of the total sample size in the first component, or with MAF  $< 0.1\%$ , were excluded from the European ancestry meta-analysis. We aggregated X-chromosome association summary statistics from the UK Biobank (19,423 SNPs) with corresponding statistics from the European GWAS component using fixed effects  $P$ -value-based meta-analysis in METAL (ref. 40) (max  $n = 99,152$ ).

We were concerned that self-reported BW as adults in the UK Biobank would not be comparable with that obtained from more stringent collection methods used in other European ancestry GWAS. In addition, the UK Biobank lacked information on gestational age for adjustment, which could have an impact on strength of association compared with the results obtained from other European ancestry GWAS. However, we observed no evidence of heterogeneity in BW allelic effects at lead SNPs between the two components of European ancestry meta-analysis, using Cochran's  $Q$  statistic<sup>41</sup> implemented in GWAMA (ref. 39) after Bonferroni correction ( $P > 0.00083$ ) (Supplementary Table 3). We tested for heterogeneity in allelic effects between studies within the European component using Cochran's  $Q$ . At loci demonstrating evidence of heterogeneity, we confirmed that association signals were not driven by outlying studies by visual inspection of forest plots. We performed sensitivity analyses to assess the impact of covariate adjustment (gestational age and population structure) on heterogeneity.

We were also concerned that overlap of individuals (duplicated or related) between the two components of the European ancestry meta-analysis might lead to false positive association signals. We performed bivariate linkage-disequilibrium score regression<sup>8</sup> using the two components of the European ancestry meta-analysis and observed a genetic covariance intercept of  $0.0156 \pm 0.0058$  (s.e.), indicating a maximum of 1,119 duplicate individuals. Univariate linkage-disequilibrium score regression<sup>8</sup> of the European ancestry meta-analysis estimated the intercept as 1.0426, which may indicate population structure or relatedness that was not adequately accounted for in the analysis. To assess the impact of this inflation on the European ancestry meta-analysis, we expanded the standard errors of BW allelic effect size estimates and re-calculated association  $P$  values. On the basis of this adjusted analysis, only the lead SNP at *MTNR1B* dropped below genome-wide significance ( $rs10830963$ ,  $P = 5.5 \times 10^{-8}$ ).

**Trans-ancestry meta-analysis.** The trans-ancestry meta-analysis combined the two European ancestry components with an additional 10,104 individuals from six GWAS from diverse ancestry groups: African American, Chinese, Filipino, Surinamese, Turkish and Moroccan. Within each GWAS, we first combined sex-specific BW association summary statistics in a fixed-effects meta-analysis, implemented in GWAMA (ref. 39) and applied a second round of genomic control<sup>34</sup>. Subsequently, we combined association summary statistics from the six non-European GWAS and the two European ancestry components in a trans-ancestry fixed-effects meta-analysis, implemented in GWAMA (ref. 39). Variants failing GWAS quality control filters in the UK Biobank, reported in less than 50% of the total sample size in the first component, or with MAF  $< 0.1\%$ , were excluded from the trans-ancestry meta-analysis. We tested for heterogeneity in allelic effects between ancestries using Cochran's  $Q$  (ref. 41).

**Approximate conditional analysis.** We searched for multiple distinct BW association signals in each of the established and novel loci, defined as 1 Mb up- and down-stream of the lead SNP from the trans-ancestry meta-analysis, through approximate conditional analysis. We applied GCTA (ref. 42) to identify 'index SNPs' for distinct association signals attaining genome-wide significance ( $P < 5 \times 10^{-8}$ ) in the European ancestry meta-analysis using a reference sample of 5,000 individuals of white British origin, randomly selected from the UK Biobank, to approximate patterns of linkage disequilibrium between variants in these regions. Note that we performed approximate conditioning on the basis of only the European ancestry meta-analysis because GCTA cannot accommodate linkage-disequilibrium variation between diverse populations.

**Prioritizing candidate genes in each BW locus.** We combined a number of approaches to prioritize the most likely candidate gene(s) in each BW locus. Expression quantitative trait loci (eQTLs) were obtained from the Genotype Tissue Expression (GTEx) Project<sup>43</sup>, the GEUVADIS project<sup>44</sup> and eleven other studies<sup>45-55</sup> using HaploReg v4 (ref. 56). We interrogated coding variants for each BW lead SNP and its proxies ( $EUR R^2 > 0.8$ ) using Ensembl<sup>57</sup> and HaploReg. Their likely functional consequences were predicted by SIFT (ref. 58) and PolyPhen2 (ref. 59). Biological candidacy was assessed by presence in significantly enriched gene set pathways from MAGENTA analyses (see below for details). We extracted all genes within 300 kb of all lead BW SNPs and searched for connectivity between



any genes using STRING (ref. 60). If two or more genes between two separate BW loci were connected, they were given an increased prior for both being plausible candidates. We also applied protein–protein interaction (PPI) analysis (see below for details) to all genes within 300 kb of each lead BW SNPs and ranked the genes based on the score for connectivity with the surrounding genes.

**Evaluation of imputation quality of the low-frequency variant at the *YKT6–GCK* locus.** At the *YKT6–GCK* locus, the lead SNP (rs138715366) was found at a low frequency in European ancestry populations (MAF = 0.92%) and was even rarer in other ancestry groups (MAF = 0.23% in African Americans, otherwise monomorphic) and was not present in the HapMap reference panel<sup>61</sup>. To assess the accuracy of imputation for this low-frequency variant, we genotyped rs138715366 in the Northern Finland Birth Cohort (NFBC) 1966 (Supplementary Table 1). Of the 5,009 samples in the study, 4,704 were successfully imputed and genotyped (or sequenced) for rs138715366. The overall concordance rate between imputed and directly assayed genotypes was 99.8% and for directly assayed heterozygote calls was 75.0%.

**Fine-mapping analyses.** We investigated linkage-disequilibrium differences between populations contributing to the trans-ancestry meta-analysis and to take advantage of the improved coverage of common and low-frequency variation offered by 1000G or 1000G and UK10K combined imputation to localize variants driving each distinct association signal achieving locus-wide significance. For each distinct signal, we used MANTRA (ref. 62) to construct 99% credible sets of variants<sup>63</sup> that together account for 99% of the posterior probability of driving the association. MANTRA incorporates a prior model of relatedness between studies, based on mean pair-wise allele frequency differences across loci, to account for heterogeneity in allelic effects (Supplementary Table 3). MANTRA has been demonstrated, by simulation, to improve localization of causal variants compared with either a fixed- or random-effects trans-ancestry meta-analysis<sup>62,64</sup>.

For loci with only one signal of association, we used MANTRA to combine summary statistics from the six non-European GWAS and the two European ancestry components. However, for loci with multiple distinct association signals, we used MANTRA to combine summary statistics from approximate conditioning for the two European components, separately for each signal.

For each distinct signal, we calculated the posterior probability that the  $j$ th variant,  $\pi_{Cj}$ , is driving the association, given by

$$\pi_{Cj} = \frac{A_j}{\sum_k A_k}$$

where the summation is over all variants mapping within the (conditional) meta-analysis across the locus. In this expression,  $A_j$  is the Bayes' factor in favour of association from the MANTRA analysis. A 99% credible set<sup>63</sup> was then constructed by: (i) ranking all variants according to their Bayes' factor,  $A_j$ ; and (ii) including ranked variants until their cumulative posterior probability exceeds 0.99.

**Genomic annotation.** We used genomic annotations of DNaseI hypersensitive sites (DHS) from the ENCODE (ref. 65 project and protein coding genes from GENCODE (ref. 66)). We filtered cell types that are cancer cell lines (karyotype 'cancer' from <https://genome.ucsc.edu/ENCODE/cellTypes.html>), and merged data from multiple samples from the same cell type. This resulted in 128 DHS cell-type annotations, as well as 4 gene-based annotations (coding exon, 5'UTR, 3'UTR and 1 kb upstream of the transcription start site (TSS)). First, we tested for the effect of each cell type DHS and gene annotation individually using the Bayes' factors for all variants in the 62 credible sets using fgwas (ref. 67). Second, we categorized the annotations into 'genetic', 'fetal DHS', 'embryonic DHS', 'stem cell DHS', 'neonatal DHS' and 'adult DHS' based on the description fields from ENCODE, and tested for the effect of each category individually as described above using fgwas. Third, we then tested the effect of each category by including all categories in a joint model using fgwas. For each of the three analyses, we obtained the estimated effects and 95% confidence intervals (CI) for each annotation, and considered an annotation enriched if the 95% CI did not overlap zero.

**Estimation of genetic variance explained.** The 'variance explained' statistic was calculated using the REML method implemented in GCTA (ref. 68). We considered the variance explained by two sets of SNPs: (i) lead SNPs of all 62 distinct association signals at the 59 established and novel autosomal BW loci identified in the European-specific or trans-ancestry meta-analyses; (ii) lead SNPs of 55 distinct association signals at the 52 novel autosomal BW loci (Extended Data Table 1a and Supplementary Table 7). The 'variance explained' was calculated in samples of European ancestry in the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study<sup>69</sup> (independent of the meta-analysis) and two studies that were part of the European ancestry meta-analysis: NFBC1966 and Generation R (Supplementary Table 1). In each study, the genetic relationship matrix was estimated for each set of SNPs and was tested individually against BW (males

and females combined) with study specific covariates. These analyses provided an estimate and s.e. for the variance explained by each of the given sets of SNPs.

**Examining the relative effects on BW of maternal and fetal genotype at the 60 identified loci.** We performed four sets of analyses. First, we used GWAS data from 4,382 mother–child pairs in the Avon Longitudinal Study of Parents and Children (ALSPAC) study to fit a 'maternal-GCTA model'<sup>6</sup> to estimate the extent to which the maternal genome might influence offspring BW independent of the fetal genome. The maternal-GCTA model uses genome-wide genetic similarity between mothers and offspring to partition the phenotypic variance in BW into components due to the maternal genotype, the child's genotype, the covariance between the two and environmental sources of variation.

Second, we compared associations with BW of the fetal versus maternal genotype at each of the 60 BW loci. The maternal allelic effect on offspring BW was obtained from a maternal GWAS meta-analysis of 68,254 European mothers from the EGG Consortium ( $n = 19,626$ )<sup>7</sup> and the UK Biobank ( $n = 48,628$ ). In the UK Biobank, mothers were asked to report the BW of their first child. Women of European ancestry with genotype data available in the May 2015 data release were included, and those with reported BW equivalent to <2.5 kg or >4.5 kg were excluded. No information on gestational age or gender of child was available. BW of first child was associated with maternal factors such as smoking status, BMI and height in the expected directions. Of the 68,254 women included in the maternal GWAS, 13% were mothers of individuals included in the current fetal European ancestry GWAS, and a further ~45% were themselves (with their own BW) included in the fetal GWAS.

Third, we additionally conducted analyses in 12,909 mother–child pairs from nine contributing studies: at each of the 60 loci, we compared the effect of the fetal genotype on BW adjusted for sex and gestational age, with and without adjustment for maternal genotype. We reciprocally compared the association between the maternal genotype and BW with and without adjustment for fetal genotype.

Fourth, we used the method of Zhang *et al.*<sup>15</sup> to test associations between BW and the maternal untransmitted, maternal transmitted and inferred paternal transmitted haplotype score of 422 height SNPs<sup>25</sup>, 30 BP SNPs<sup>13,14</sup> and 84 T2D SNPs<sup>24</sup> in 5,201 mother–child pairs from the ALSPAC study.

**Linkage-disequilibrium score regression.** The use of linkage-disequilibrium score regression to estimate the genetic correlation between two traits/diseases has been described in detail elsewhere<sup>70</sup>. Briefly, the linkage-disequilibrium score is a measure of how much genetic variation each variant tags; if a variant has a high linkage-disequilibrium score then it is in high linkage disequilibrium with many nearby polymorphisms. Variants with high linkage-disequilibrium scores are more likely to contain more true signals and hence provide more chance of overlap with genuine signals between GWAS. The linkage-disequilibrium score regression method uses summary statistics from the GWAS meta-analysis of BW and the other traits of interest, calculates the cross-product of test statistics at each SNP, and then regresses the cross-product on the linkage-disequilibrium score. Bulik-Sullivan *et al.*<sup>70</sup> show that the slope of the regression is a function of the genetic covariance between traits:

$$E(z_1 z_2) = \frac{\sqrt{N_1 N_2} \rho_g}{M} l_j + \frac{\rho_{N_s}}{\sqrt{N_1 N_2}}$$

where  $N_i$  is the sample size for study  $i$ ,  $\rho_g$  is the genetic covariance,  $M$  is the number of SNPs in the reference panel with MAF between 5% and 50%,  $l_j$  is the linkage-disequilibrium score for SNP  $j$ ,  $N_s$  quantifies the number of individuals that overlap both studies, and  $\rho$  is the phenotypic correlation amongst the  $N_s$  overlapping samples. Thus, if there is sample overlap (or cryptic relatedness between samples), it will only affect the intercept from the regression (that is, the term  $\frac{\rho_{N_s}}{\sqrt{N_1 N_2}}$ ) and not the slope, and hence estimates of the genetic covariance will not be biased by sample overlap. Likewise, population stratification will affect the intercept but will have minimal impact on the slope (that is, intuitively since population stratification does not correlate with linkage disequilibrium between nearby markers).

Summary statistics from the GWAS meta-analysis for traits and diseases of interest were downloaded from the relevant consortium website. The summary statistics files were reformatted for linkage-disequilibrium score regression analysis using the `munge_sumstats.py` python script provided on the developer's website (<https://github.com/bulik/ldsc>). For each trait, we filtered the summary statistics to the subset of HapMap 3 SNPs<sup>71</sup>, as advised by the developers, to ensure that no bias was introduced due to poor imputation quality. Summary statistics from the European-specific BW meta-analysis were used because of the variable linkage-disequilibrium structure between ancestry groups. Where the sample size for each SNP was included in the results file this was flagged using N-col; if no sample size was available then the maximum sample size reported in the reference for the GWAS meta-analysis was used. SNPs were excluded for the

following reasons:  $MAF < 0.01$ ; ambiguous strand; duplicate rsID; non-autosomal SNPs; reported sample size less than 60% of the total available. Once all files were reformatted, we used the `ldsc.py` python script, also on the developers' website, to calculate the genetic correlation between BW and each of the traits and diseases. The European linkage-disequilibrium score files calculated from the 1000G reference panel and provided by the developers were used for the analysis. Where multiple GWAS meta-analyses had been conducted on the same phenotype (that is, over a period of years), the genetic correlation with BW was estimated using each set of summary statistics and presented in Supplementary Table 12. The phenotypes with multiple GWAS included height, BMI, waist-hip ratio (adjusted for BMI), total cholesterol, triglycerides, high density lipoprotein (HDL) and low density lipoprotein (LDL). The estimate of the genetic correlation between the multiple GWAS meta-analyses on the same phenotype were comparable and the later GWAS had a smaller standard error due to the increased sample size, so only the genetic correlation between BW and the most recent meta-analyses were presented in Fig. 2.

In the published GWAS for blood pressure<sup>14</sup> the phenotype was adjusted for BMI. Caution is needed when interpreting the genetic correlation between BW and BMI-adjusted SBP owing to the potential for collider bias<sup>72</sup>. Since BMI is associated with both blood pressure and BW, it is possible that the use of a blood pressure genetic score adjusted for BMI might bias the genetic correlation estimate towards a more negative value. To verify that the inverse genetic correlation with BW ( $r_g = -0.26$ ,  $s.e. = 0.05$ ,  $P = 6.5 \times 10^{-9}$ ) was not due to collider bias caused by the BMI adjustment of the phenotype, we obtained an alternative estimate using UK Biobank GWAS data for SBP that was unadjusted for BMI and obtained a similar result ( $R_g = -0.22$ ,  $s.e. = 0.03$ ,  $P = 5.5 \times 10^{-13}$ ). The SBP phenotype in the UK Biobank was prepared as follows. Two blood pressure readings were taken at assessment, approximately 5 min apart. We included all individuals with an automated blood pressure reading (taken using an automated Omron blood pressure monitor). Two valid measurements were available for most participants (averaged to create a blood pressure variable, or alternatively a single reading was used if only one was available). Individuals were excluded if the two readings differed by more than 4.56 s.d. Blood pressure measurements more than 4.56 s.d. away from the mean were excluded. We accounted for blood pressure medication use by adding 15 mm Hg to the SBP measure. Blood pressure was adjusted for age, sex and centre location and then inverse rank normalized. We performed the GWAS on 127,698 individuals of British descent using BOLT-LMM (ref. 37), with genotyping array as covariate.

**Estimating the proportion of the BW-adult traits covariance attributable to genotyped SNPs.** We estimated the phenotypic, genetic and residual correlations as well as the genetic and residual covariance between BW and several quantitative traits and/or disease outcomes in the UK Biobank using directly genotyped SNPs and the REML method implemented in BOLT-LMM (ref. 37). The traits examined included T2D, SBP, diastolic blood pressure, CAD, height, BMI, weight, waist-hip ratio, hip circumference, waist circumference, obesity, overweight, age at menarche, asthma, and smoking. Where phenotypes were not available (for example, serum blood measures are not currently available in the UK Biobank), we obtained estimates using the NFBC1966 study (for correlations/covariance between BW and triglycerides, total cholesterol, HDL, LDL, fasting glucose and fasting insulin). In the UK Biobank analysis, we used 57,715 unrelated individuals with BW available and identified by the UK Biobank as white British. SNPs with evidence of deviation from Hardy-Weinberg equilibrium ( $P < 1 \times 10^{-6}$ ),  $MAF < 0.05$  or overall missing rate  $> 0.015$  were excluded, resulting in 328,928 SNPs for analysis. We included the first five ancestry principal components as covariates. In the NFBC1966 analysis, 5,009 individuals with BW were enrolled. Genotyped SNPs that passed quality control (Supplementary Table 2) were included, resulting in 324,895 SNPs for analysis. The first three ancestry principal components and sex were included as covariates.

**Gene set enrichment analysis.** Meta-analysis gene-set enrichment of variant associations (MAGENTA) was used to explore pathway-based associations using summary statistics from the trans-ancestry meta-analysis. MAGENTA implements a gene set enrichment analysis (GSEA) based approach, as previously described<sup>9</sup>. Briefly, each gene in the genome was mapped to a single index SNP with the lowest  $P$  value within a 110 kb upstream and 40 kb downstream window. This  $P$  value, representing a gene score, was then corrected for confounding factors such as gene size, SNP density and linkage-disequilibrium-related properties in a regression model. Genes within the HLA-region were excluded from analysis due to difficulties in accounting for gene density and linkage-disequilibrium patterns. Each mapped gene in the genome was then ranked by its adjusted gene score. At a given significance threshold (95th and 75th percentiles of all gene scores), the observed number of gene scores in a given pathway, with a ranked score above the specified threshold percentile, was calculated. This observed statistic was

then compared to 1,000,000 randomly permuted pathways of identical size. This generates an empirical GSEA  $P$  value for each pathway. Significance was attained when an individual pathway reached a  $FDR < 0.05$  in either analysis. In total, 3,216 pre-defined biological pathways from Gene Ontology, PANTHER, KEGG and Ingenuity were tested for enrichment of multiple modest associations with BW. The MAGENTA software was also used for enrichment testing of custom gene sets.

**Protein-protein interaction network analyses.** We used the integrative protein-interaction-network-based pathway analysis (iPINBPA) method<sup>73</sup>. Briefly, we generated gene-wise  $P$  values from the trans-ancestry meta-analysis using VEGAS2 (ref. 74), which mapped the SNPs to genes and accounted for possible confounders, such as linkage-disequilibrium between markers. The empirical gene-wise  $P$  values were calculated using simulations from the multivariate normal distribution. Those that were nominally significant ( $P \leq 0.01$ ) were selected as 'seed genes', and were collated within a high confidence version of inweb3 (ref. 75) to weight the nodes in the network following a guilt-by-association approach. In a second step, a network score was defined by the combination of the  $Z$  scores derived from the gene-wise  $P$  values with node weights using the Liptak-Stouffer method<sup>76</sup>. A heuristic algorithm was then applied to extensively search for modules enriched in genes with low  $P$  values. The modules were further normalized using a null distribution of 10,000 random networks. Only those modules with  $Z$  score  $> 5$  were selected. Finally, the union of all modules constructed a BW-overall PPI network. Both the proteins on the individual modules and on the overall BW-PPI were interrogated for enrichment in Gene Ontology terms (biological processes) using a hypergeometric test. Terms were considered as significant when the adjusted  $P$  value, following the Benjamini-Hochberg procedure, was below 0.05.

**Point of contact analyses.** The same methodology described above was applied to 16 different adult traits resulting in a number of enriched modules per trait. Different modules for each trait were combined in a single component and the intersection between these trait-specific components and the BW component was calculated. This intersection was defined as the PoC network. We used the resulting PoC networks in downstream analyses to interrogate which set of proteins connected BW variation and adult trait variation via pathways enriched in the overall BW analysis.

**Parent-of-origin specific associations.** We first searched for evidence of parent-of-origin effects in the UK Biobank samples by comparing variance between heterozygotes and homozygotes using Quicktest (ref. 77). In this analysis, we used only unrelated individuals identified genetically as of white British origin ( $n = 57,715$ ). Principal components were generated using these individuals and the first five were used to adjust for population structure as covariates in the analysis, in addition to a binary indicator for genotyping array.

We also examined 4,908 mother-child pairs in ALSPAC and determined the parental origin of the alleles where possible<sup>78</sup>. Briefly, the method used mother-child pairs to determine the parent of origin of each allele. For example, if the mother/child genotypes were AA/Aa, the child's maternal/paternal allele combination was A/a. For the situation where both mother and child were heterozygous, the child's maternal/paternal alleles could not be directly specified. However, the parental origin of the alleles could be determined by phasing the genotype data and comparing maternal and child haplotypes. We then tested these alleles for association with BW adjusting for sex and gestational age.

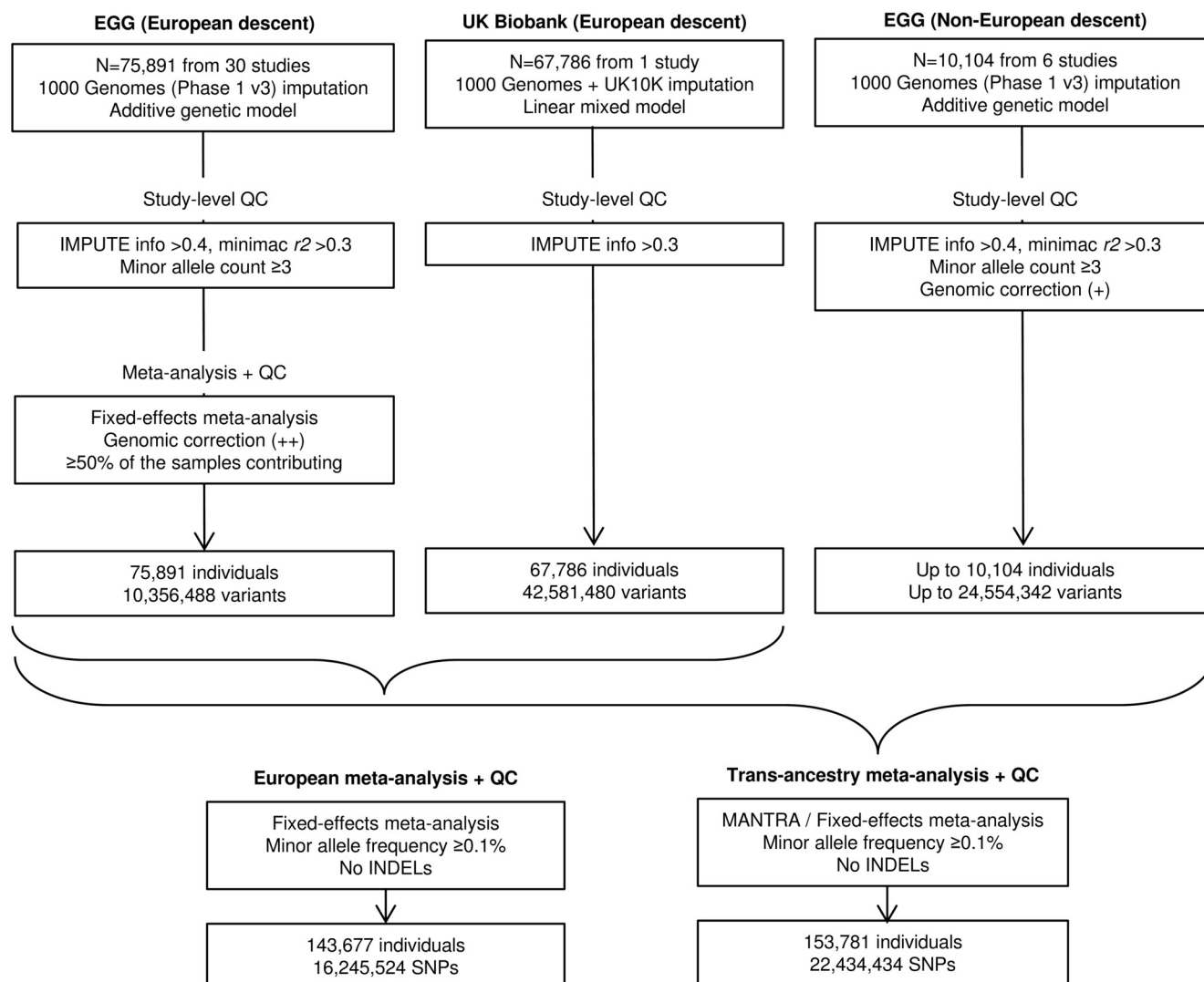
Statistical power in these currently available sample sizes was insufficient to rule out widespread parent-of-origin effects across the regions tested. Using the mean  $\beta$  (0.034 s.d.) and  $MAF$  (0.28) of the identified loci, we estimate that we would need at least 200,000 unrelated individuals or 70,000 mother-child pairs for 80% power to detect parent-of-origin effects at  $P < 0.00085$ .

**Hierarchical clustering of BW loci.** To explore the different patterns of association between BW and other anthropometric/metabolic/endocrine traits and diseases, we performed hierarchical clustering analysis. The lead SNP (or proxy,  $R^2 > 0.6$ ) at the 60 BW loci was queried in publicly available GWAS meta-analysis datasets or in GWAS results obtained through collaboration<sup>79</sup>. Results were available for 53 of those loci and the extracted  $Z$  score (allelic effect/s.e., Supplementary Table 17) was aligned to the BW-raising allele. We performed two dimensional clustering by trait and by locus. We computed the Euclidean distance amongst  $Z$  scores of the extracted traits and loci and performed complete hierarchical clustering implemented in the `pvcust` package (<http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvcust/>) in R v3.2.0 (<http://www.R-project.org/>). Clustering uncertainty was measured by multiscale bootstrap resampling estimated from 1,000 replicates. We used  $\alpha = 0.05$  to define distinct clusters and, based on the bootstrap analysis, calculated the Calinski index to identify the number of well-supported clusters (`cascadeKM` function, `vegan` package, <http://CRAN.R-project.org/package=vegan>). Clustering was visualized by constructing dendrograms and a heat map.

Separately from the hierarchical clustering analysis, we queried the lead SNP at *EPAS1* in a GWAS of haematological traits<sup>80</sup> because variation at that locus has previously been implicated in BW and adaptation to hypoxia at high altitudes in Tibetans<sup>81,82</sup> (Supplementary Table 17).

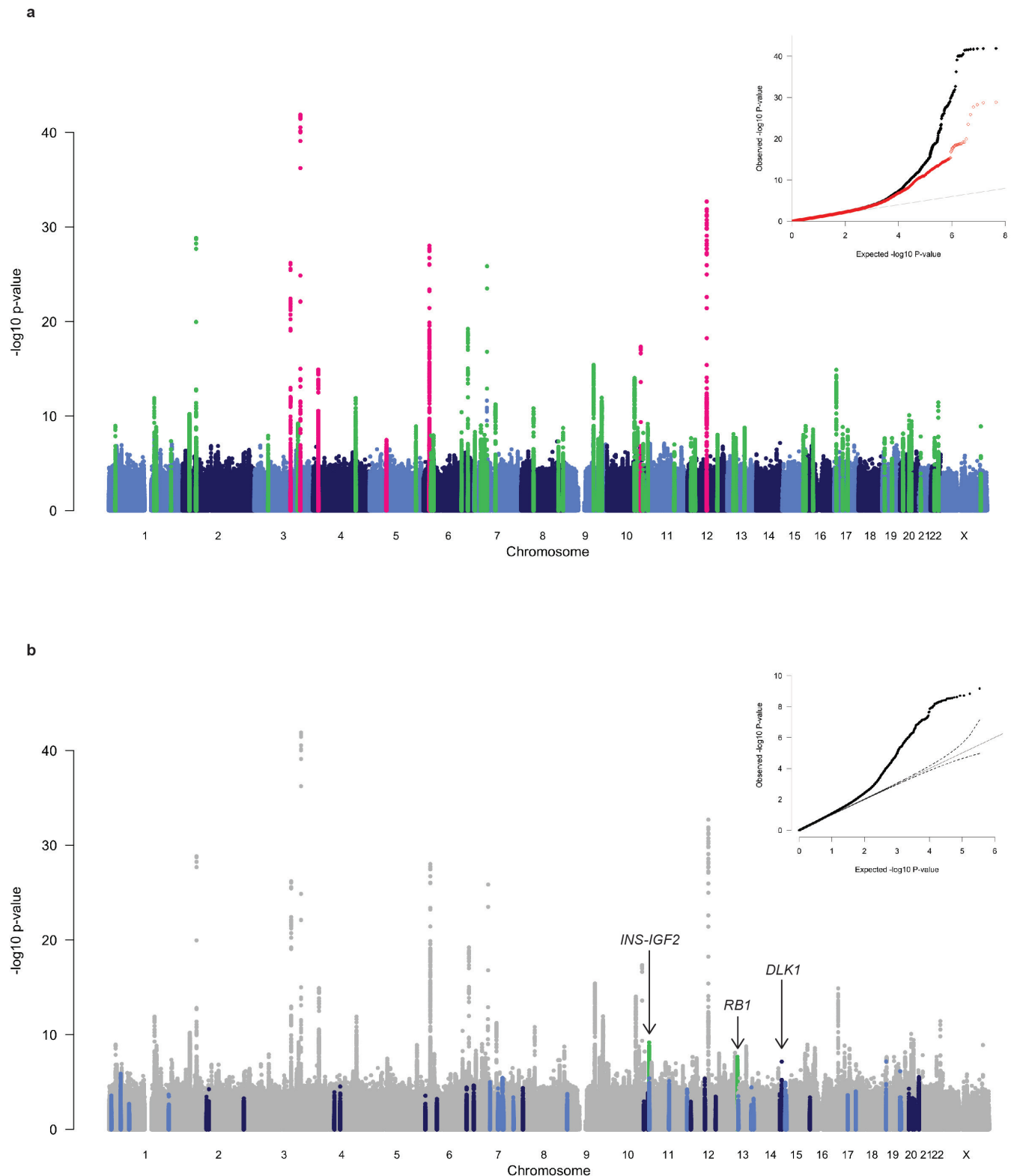
30. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
31. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
32. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
33. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
34. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
35. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
36. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK Biobank data: come and get it. *Sci. Transl. Med.* **6**, 224ed4 (2014).
37. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
38. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
39. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11**, 288 (2010).
40. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
41. Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* **2**, e841 (2007).
42. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375, S1–S3 (2012).
43. GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
44. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
45. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773–777 (2010).
46. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
47. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
48. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
49. Li, Q. *et al.* Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. *Hum. Mol. Genet.* **23**, 5294–5302 (2014).
50. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
51. Zou, F. *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet.* **8**, e1002707 (2012).
52. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8**, e1003029 (2012).
53. Koopmann, T. T. *et al.* Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. *PLoS One* **9**, e97380 (2014).
54. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
55. Grundberg, E. *et al.* Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. *PLoS Genet.* **7**, e1001279 (2011).
56. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
57. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749–D755 (2014).
58. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
59. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
60. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
61. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
62. Morris, A. P. Transethnic meta-analysis of genome-wide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).
63. The Wellcome Trust Case Control Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
64. Wang, X. *et al.* Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **22**, 2303–2311 (2013).
65. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
66. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
67. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
68. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
69. Urbanek, M. *et al.* The chromosome 3q25 genomic region is associated with measures of adiposity in newborns in a multi-ethnic genome-wide association study. *Hum. Mol. Genet.* **22**, 3583–3596 (2013).
70. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
71. The International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
72. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96**, 329–339 (2015).
73. Wang, L., Mousavi, P. & Baranzini, S. E. iPINBPA: an integrative network-based functional module discovery tool for genome-wide association studies. *Pac. Symp. Biocomput.* 255–266 (2015).
74. Mishra, A. & Macgregor, S. VEGAS2: software for more flexible gene-based testing. *Twin Res. Hum. Genet.* **18**, 86–91 (2015).
75. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316 (2007).
76. Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18**, 1368–1373 (2005).
77. Hoggart, C. J. *et al.* Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index. *PLoS Genet.* **10**, e1004508 (2014).
78. Wang, S., Yu, Z., Miller, R. L., Tang, D. & Perera, F. P. Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. *Hum. Hered.* **71**, 196–208 (2011).
79. Painter, J. N. *et al.* Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat. Genet.* **43**, 51–54 (2011).
80. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* **41**, 1191–1198 (2009).
81. Xu, X. H. *et al.* Two functional loci in the promoter of *EPAS1* gene involved in high-altitude adaptation of Tibetans. *Sci. Rep.* **4**, 7465 (2014).
82. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).





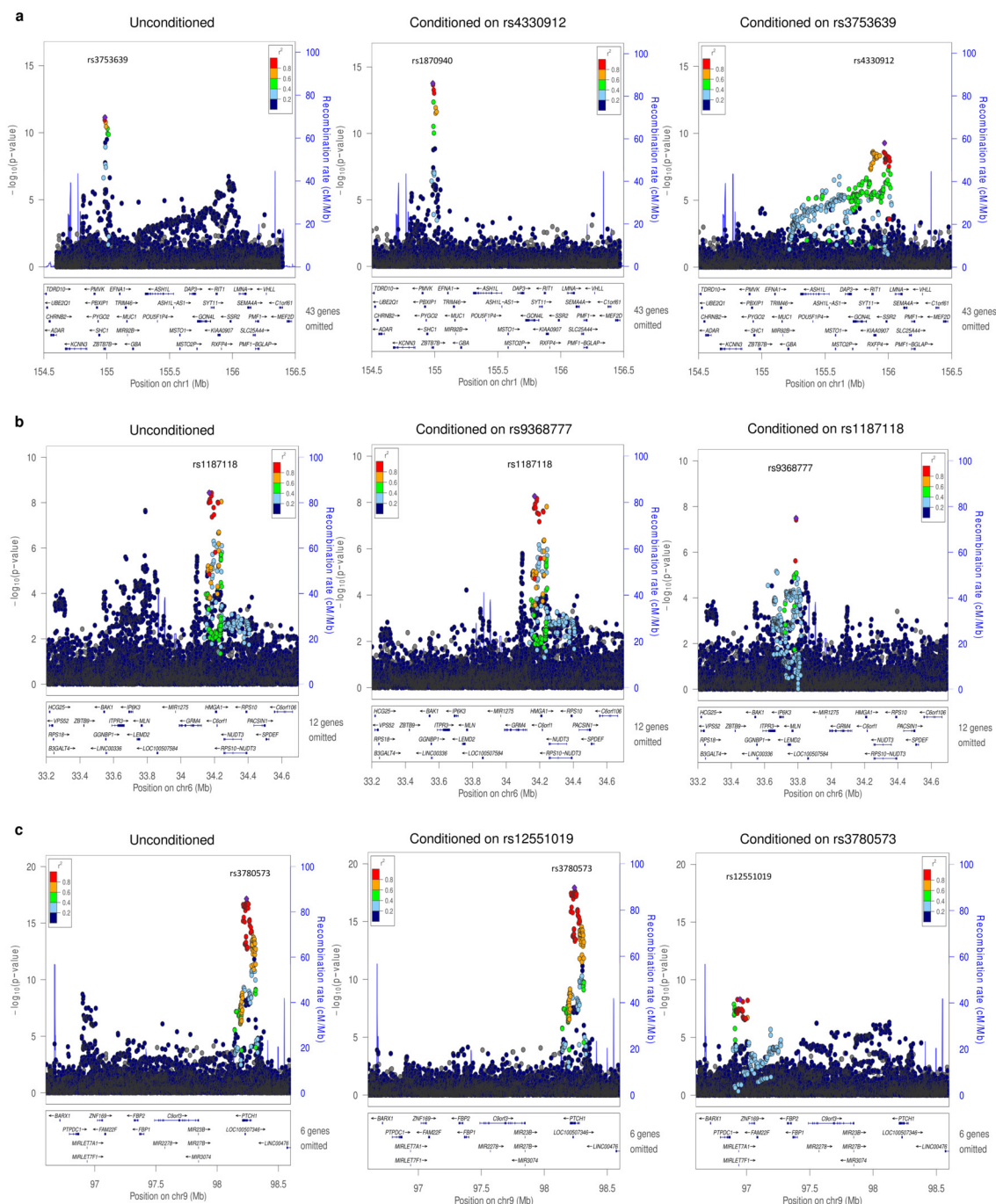
Extended Data Figure 1 | Flow chart of the study design.





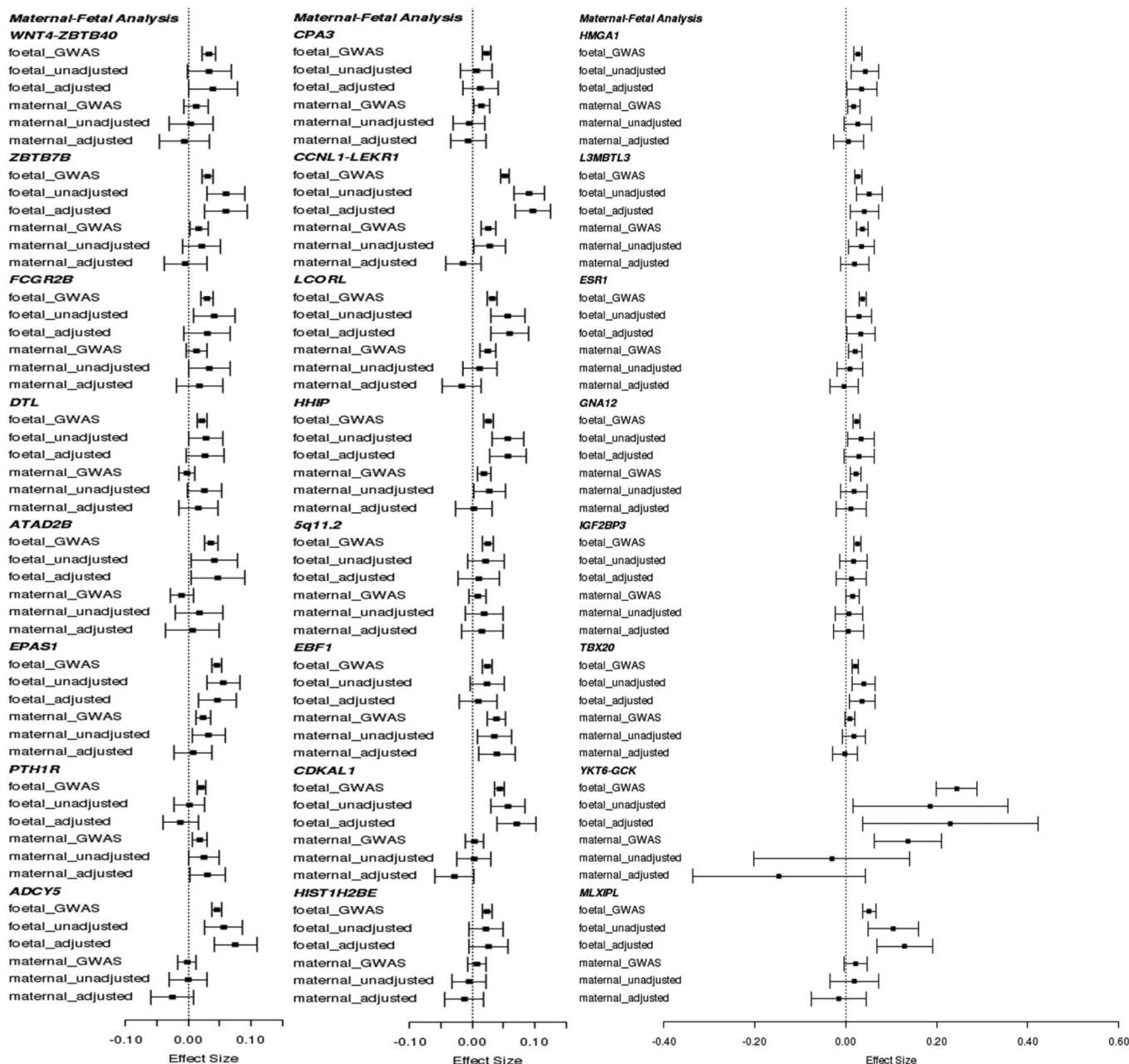
**Extended Data Figure 2 | Manhattan and quantile–quantile (QQ) plots of the trans-ancestry meta-analysis for BW.** **a**, Manhattan (main panel) and QQ (top right) plots of genome-wide association results for BW from trans-ancestry meta-analysis of up to 153,781 individuals. The association  $P$  value (on  $-\log_{10}$  scale) for each of up to 22,434,434 SNPs ( $y$  axis) was plotted against the genomic position (NCBI Build 37;  $x$  axis). Association signals that reached genome-wide significance ( $P < 5 \times 10^{-8}$ ) are shown in green if novel and pink if previously reported. In the QQ plot, the black dots represent observed  $P$  values and the grey line represents expected  $P$  values under the null distribution. The red dots represent observed  $P$  values after excluding the previously identified signals<sup>5</sup>. **b**, Manhattan

(main panel) and QQ (top right) plots of trans-ethnic GWAS meta-analysis for BW highlighting the reported imprinted regions described in Supplementary Table 14. Novel association signals that reached genome-wide significance ( $P < 5 \times 10^{-8}$ ) and mapped to imprinted regions are shown in green. Genomic regions outside imprinted regions are shaded in grey. SNPs in the imprinted regions are shown in light blue or dark blue, depending on chromosome number (odd or even). In the QQ plot, the black dots represent observed  $P$  values and the grey lines represent expected  $P$  values and their 95% confidence intervals under the null distribution for the SNPs within the imprinted regions.



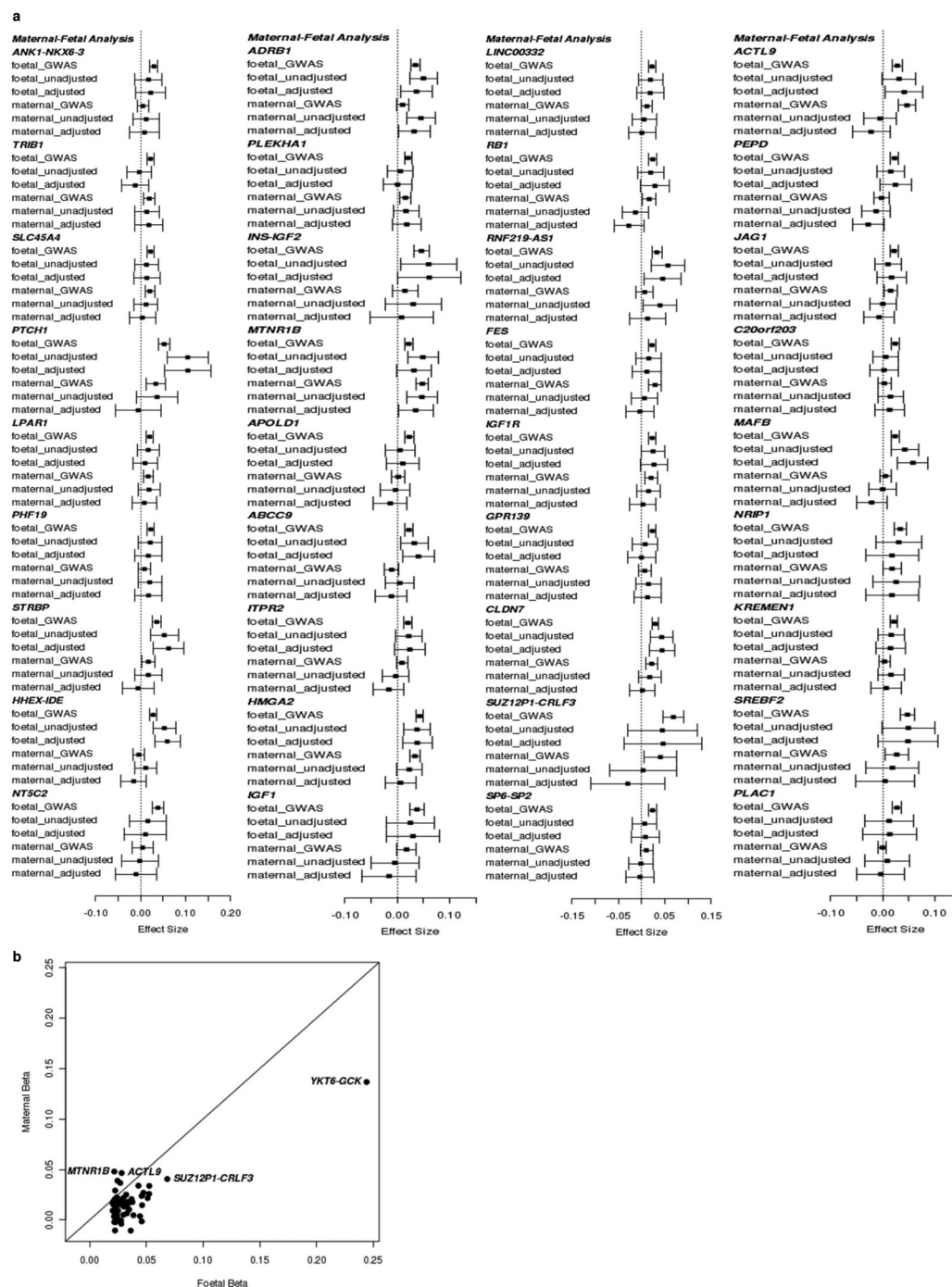
**Extended Data Figure 3 | Regional plots for multiple distinct signals at three BW loci.** Regional plots for each locus, *ZBTB7B* (a), *HMGA1* (b) and *PTCH1* (c), are displayed from: the unconditional European-specific meta-analysis of up to 143,677 individuals (left); the approximate conditional meta-analysis for the primary signal after adjustment for the index variant for the secondary signal (middle); and the approximate conditional meta-analysis for the secondary signal after adjustment for the index variant for the primary signal (right). Directly genotyped or imputed

SNPs were plotted with their association  $P$  values (on a  $-\log_{10}$  scale) as a function of genomic position (NCBI Build 37). Estimated recombination rates (blue lines) were plotted to reflect the local linkage-disequilibrium structure around the index SNPs and their correlated proxies. SNPs were coloured in reference to linkage-disequilibrium with the particular index SNP according to a blue to red scale from  $R^2 = 0$  to 1, based on pairwise  $R^2$  values estimated from a reference of 5,000 individuals of white British origin, randomly selected from the UK Biobank.



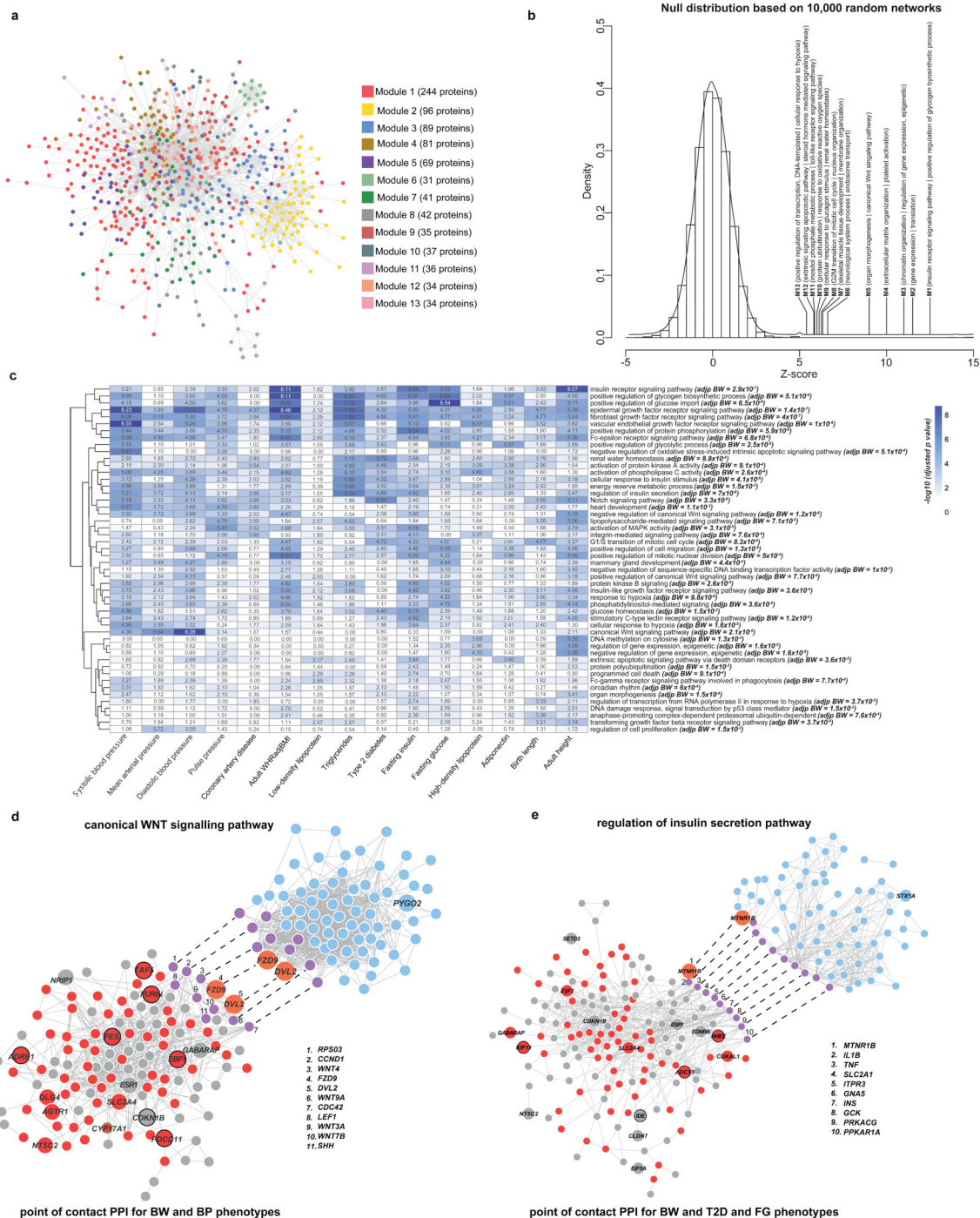
**Extended Data Figure 4 | Comparison of fetal effect sizes and maternal effect sizes at 60 known and novel birth weight loci, for the first 24 loci.** The remaining loci are shown in Extended Data Fig. 5a. For each BW locus, the following six effect sizes (with 95% CI) are shown, all aligned to the same BW-raising allele: fetal\_GWAS, fetal allelic effect on BW (from European ancestry meta-analysis of up to  $n = 143,677$  individuals); fetal\_unadjusted, fetal allelic effect on BW (unconditioned in  $n = 12,909$  mother-child pairs); fetal\_adjusted, fetal effect (conditioned on maternal genotype,  $n = 12,909$ ); maternal\_GWAS, maternal allelic effect on offspring BW (from meta-analysis of up to  $n = 68,254$  European mothers)<sup>7</sup>; maternal\_unadjusted, maternal allelic effect on offspring

BW (unconditioned,  $n = 12,909$ ); maternal\_adjusted, maternal effect (conditioned on fetal genotype,  $n = 12,909$ ). The 60 BW loci were ordered by chromosome and position (Supplementary Tables 10, 11). These plots illustrate that, in large GWAS of BW, fetal effect size estimates are larger than those of maternal at 55 out of 60 identified loci (binomial  $P = 1 \times 10^{-11}$ ), suggesting that most of the associations are driven by the fetal genotype. In conditional analyses that modelled the effects of both maternal and fetal genotypes ( $n = 12,909$  mother-child pairs), confidence intervals around the estimates were wide, precluding inference about the likely contribution of maternal versus fetal genotype at individual loci.



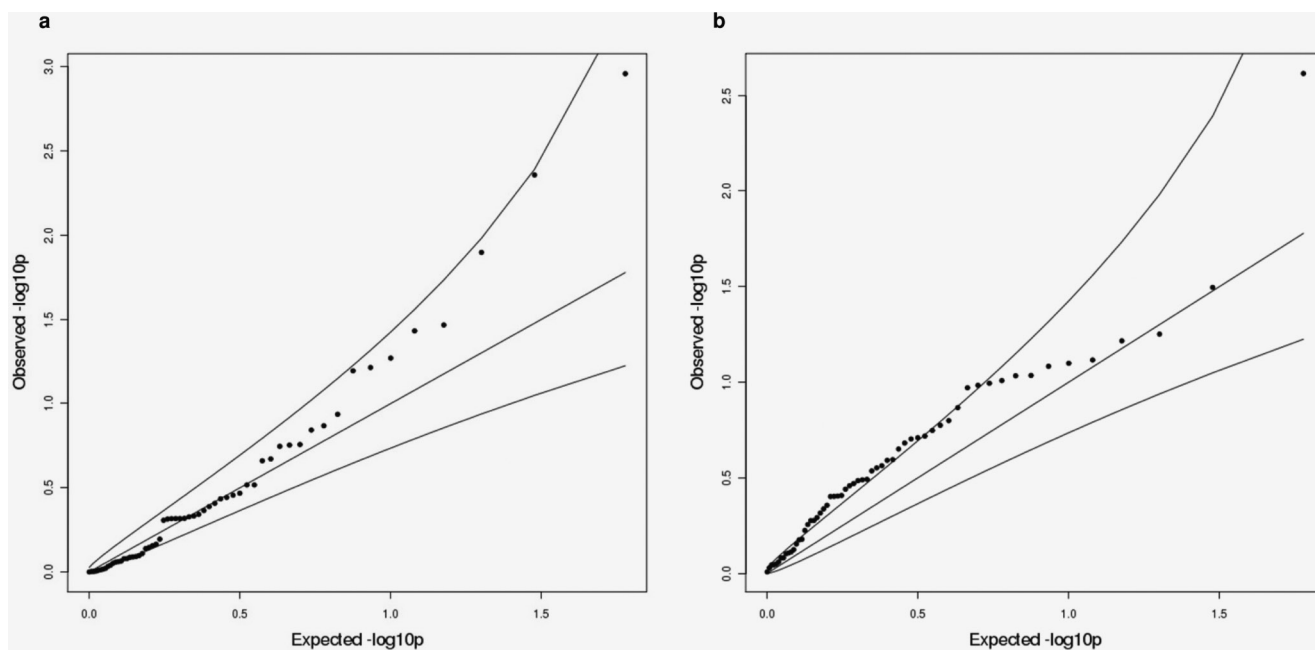
**Extended Data Figure 5 | Comparison of fetal effect sizes and maternal effect sizes at 60 known and novel birth weight loci, for the remaining 36 loci. a,** Continued from Extended Data Fig. 4. **b,** The scatter plot illustrates the difference between the fetal (x axis) and maternal (y axis) effect sizes in the overall maternal versus fetal GWAS results.





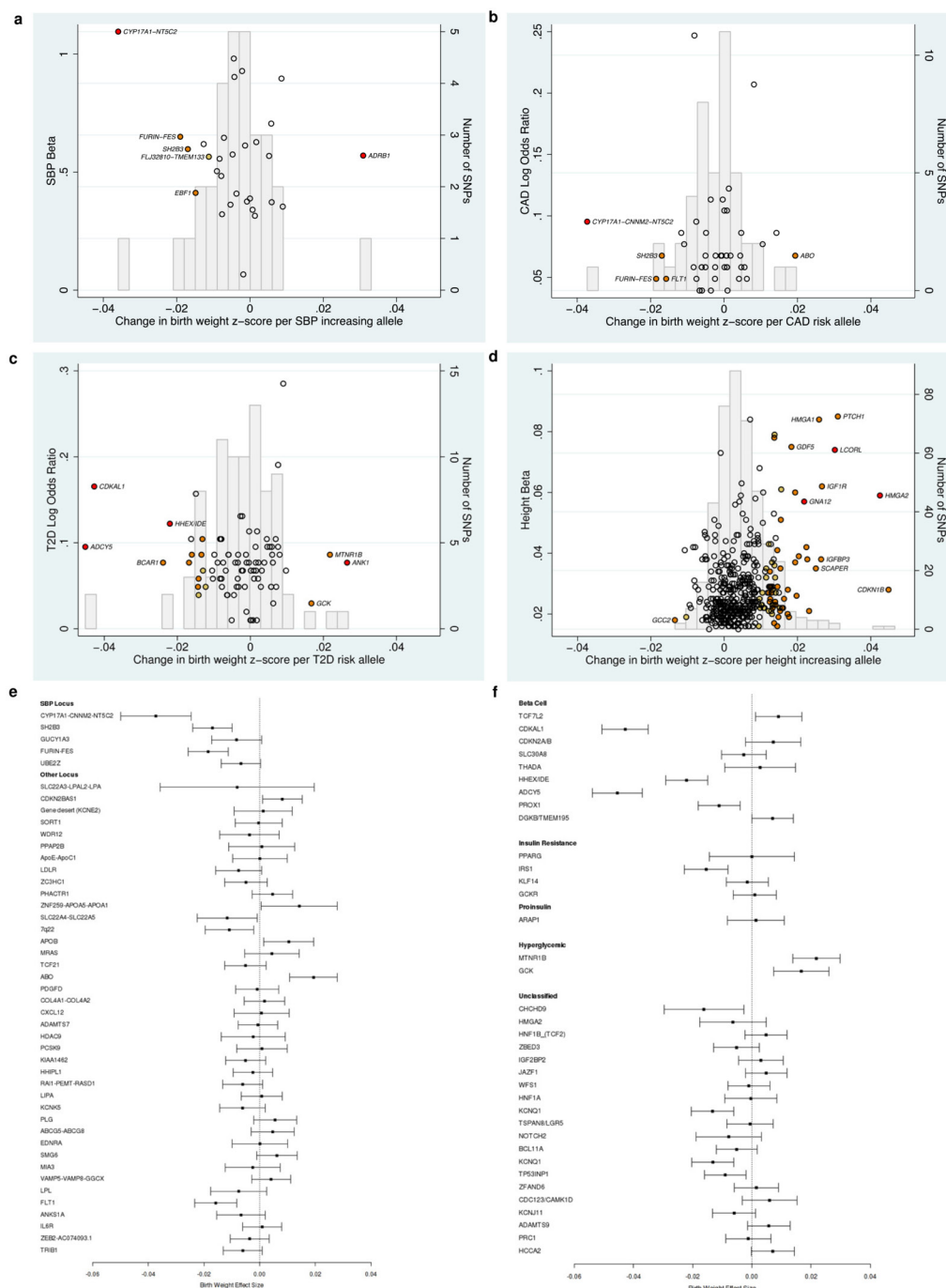
**Extended Data Figure 6 | Protein-protein Interaction (PPI) Network analysis.** **a**, The largest global component of BW PPI network containing 13 modules is shown. **b**, The histogram shows the null distribution of Z scores of BW PPI networks based on 10,000 random networks, and where the Z scores for the 13 BW modules (M1–13) lie. For each module, the two most significant GO terms are shown. **c**, A heat map is shown, which takes the top 50 biological processes over-represented in the global BW PPI network (listed at the right of the plot), and displays the extent of enrichment for the various trait-specific “point of contact” (PoC) PPI networks. **d**, **e**, Trait-specific PoC PPI networks composed of proteins that are shared in both the global BW PPI network and networks generated

using the same pipeline for each of the adult traits: **d**, canonical Wnt signalling pathway enriched for PoC PPI between BW and blood pressure (BP)-related phenotypes; and **e**, regulation of insulin secretion pathway enriched for PoC between BW and T2D/fasting glucose (FG). Red nodes indicate those present in PoC for BW and traits of interest; blue nodes correspond to the pathway nodes; purple nodes are those present in both the pathway and PoC; orange nodes are genes in BW loci that overlap with both the pathway and PoC. Large nodes correspond to genes in BW loci (within 300 kb from the lead SNP), and have a black border if they, amongst all BW loci, have a stronger (top 5) association with at least one of the pairing adult traits.



**Extended Data Figure 7 | Quantile–Quantile (QQ) plots of variance comparison between heterozygotes and homozygotes analysis in 57,715 UK Biobank samples and parent-of-origin specific analysis in 4,908 ALSPAC mother–child pairs at 59 autosomal BW loci plus *DLK1*.** **a**, QQ plot from the Quicktest analysis (ref. 77) comparing the BW variance of heterozygotes with homozygotes in 57,715 UK Biobank samples. **b**, QQ plot from the parent-of-origin specific analysis testing the association between BW and maternally transmitted versus paternally transmitted alleles in 4,908 mother–child pairs from the ALSPAC study (Methods,

Supplementary Tables 15, 16). In both panels, the black dots represent lead SNPs at 59 identified autosomal BW loci and a further sub-genome-wide significant signal for BW near *DLK1* (rs6575803;  $P = 5.6 \times 10^{-8}$ ). The grey lines represent expected  $P$  values and their 95% confidence intervals under the null distribution for the 60 SNPs. Both results show trends in favour of imprinting effects at BW loci; however, despite the large sample size, these analyses were underpowered (see Methods) and much larger sample sizes are required for definitive analysis.



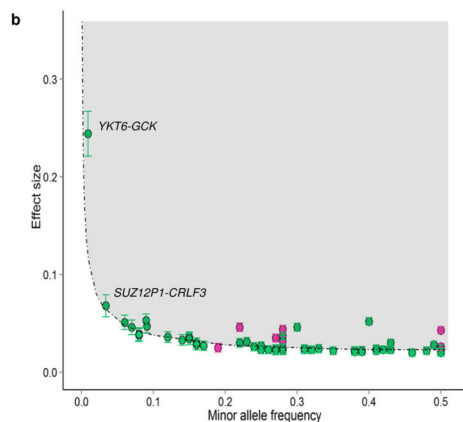
**Extended Data Figure 8 | Summary of previously reported loci for SBP, CAD, T2D and adult height and their effect on birth weight.** **a–d**, Effect sizes (left y axis) of previously reported 30 SBP loci<sup>13,14</sup>, 45 CAD loci<sup>23</sup>, 84 T2D loci<sup>24</sup> and 422 adult height loci<sup>25</sup> were plotted against effects on BW (x axis). Effect sizes were aligned to the adult trait (or risk)-raising allele. The colour of each dot indicates BW association  $P$  value: red,  $P < 5 \times 10^{-8}$ ; orange,  $5 \times 10^{-8} \leq P < 0.001$ ; yellow,  $0.001 \leq P < 0.01$ ; white,  $P \geq 0.01$ . The superimposed grey frequency histogram shows the number of SNPs (right y axis) in each category of BW effect size. **e**, Effect sizes (with 95% CI) on BW of 45 known CAD loci were plotted arranged in the order of CAD effect size from highest to lowest, separating out the known

SBP loci. CAD loci with a larger effect on BW concentrated amongst loci with primary blood pressure association. **f**, Effect sizes (with 95% CI) on BW of 32 known T2D loci were plotted, subdivided by previously reported categories derived from detailed adult physiological data<sup>27</sup>. Heterogeneity in BW effect sizes between five T2D loci groups with different mechanistic categories was substantial (Cochran's  $Q$  statistic  $P_{\text{het}} = 1.2 \times 10^{-9}$ ). In pairwise comparisons, the 'beta cell' group of variants differed from the other four groups: fasting hyperglycaemia ( $P_{\text{het}} = 3 \times 10^{-11}$ ), insulin resistance ( $P_{\text{het}} = 0.002$ ), proinsulin ( $P_{\text{het}} = 0.78$ ) and unclassified ( $P_{\text{het}} = 0.02$ ) groups. All of the BW effect sizes plotted in the forest plots were aligned to the trait (or risk)-raising allele.

**Extended Data Table 1 | Sixty loci associated with BW ( $P < 5 \times 10^{-8}$ ) in European ancestry meta-analysis of up to 143,677 individuals and/or trans-ancestry meta-analysis of up to 153,781 individuals**

**a**

Locus	Lead SNP	Chr.	Position (bp, b37)	Alleles Effect/Other	EAF	European ancestry $\beta$ (SE)	European ancestry $P$ -value	Trans-ancestry $\beta$ (SE)	Trans-ancestry $P$ -value
<b>Previously reported loci</b>									
<i>CCNL1-LEKR1</i>	rs13322435	3	156,795,468	A/G	0.59	0.053 (0.004)	$3.7 \times 10^{-41}$	0.052 (0.004)	$1.3 \times 10^{-42}$
<i>HMG2</i>	rs1351394	12	66,351,826	T/C	0.48	0.044 (0.004)	$1.9 \times 10^{-37}$	0.043 (0.004)	$2.0 \times 10^{-39}$
<i>CDKAL1</i>	rs35261542	6	20,675,792	C/A	0.73	0.044 (0.004)	$4.4 \times 10^{-37}$	0.044 (0.004)	$9.7 \times 10^{-37}$
<i>ADCY5</i>	rs11719201	3	123,068,744	T/C	0.23	0.046 (0.004)	$2.4 \times 10^{-38}$	0.046 (0.004)	$6.4 \times 10^{-27}$
<i>ADRB1</i>	rs7076936	10	115,789,375	T/C	0.73	0.036 (0.004)	$4.7 \times 10^{-18}$	0.035 (0.004)	$4.7 \times 10^{-18}$
<i>LCORL</i>	rs925098	4	17,919,811	G/A	0.28	0.034 (0.004)	$5.4 \times 10^{-16}$	0.032 (0.004)	$1.3 \times 10^{-15}$
<i>5q11.2</i>	rs854037	5	57,091,783	A/G	0.80	0.027 (0.005)	$2.2 \times 10^{-8}$	0.025 (0.005)	$3.5 \times 10^{-8}$
<b>Novel loci</b>									
<i>EPAS1</i>	rs1374204	2	46,484,205	T/C	0.70	0.047 (0.004)	$6.2 \times 10^{-39}$	0.046 (0.004)	$1.5 \times 10^{-29}$
<i>YKT6-GCK</i>	rs138715366	7	44,246,271	C/T	0.99	0.241 (0.023)	$7.2 \times 10^{-38}$	0.244 (0.023)	$1.4 \times 10^{-26}$
<i>ESR1</i>	rs1101081	6	152,032,917	C/T	0.73	0.038 (0.004)	$1.6 \times 10^{-19}$	0.037 (0.004)	$6.1 \times 10^{-20}$
<i>PTCH1</i>	rs28510415	9	98,245,026	G/A	0.09	0.056 (0.007)	$1.5 \times 10^{-17}$	0.053 (0.006)	$4.0 \times 10^{-16}$
<i>CLDN7</i>	rs113086489	17	7,171,356	T/C	0.55	0.031 (0.004)	$9.1 \times 10^{-16}$	0.030 (0.004)	$1.3 \times 10^{-15}$
<i>HHEX-IDE</i>	rs61862780	10	94,468,643	T/C	0.52	0.028 (0.004)	$3.0 \times 10^{-14}$	0.028 (0.004)	$9.5 \times 10^{-15}$
<i>STRBP</i>	rs700059	9	125,824,055	G/A	0.16	0.033 (0.005)	$4.7 \times 10^{-10}$	0.036 (0.005)	$1.2 \times 10^{-12}$
<i>HHIP</i>	rs6537307	4	145,601,863	G/A	0.48	0.025 (0.004)	$9.5 \times 10^{-12}$	0.026 (0.004)	$1.3 \times 10^{-12}$
<i>ZBTB7B</i>	rs3753639	1	154,986,091	C/T	0.23	0.031 (0.004)	$7.3 \times 10^{-12}$	0.031 (0.004)	$1.3 \times 10^{-12}$
<i>SREBF2</i>	rs62240862	22	42,259,524	C/T	0.92	0.047 (0.007)	$9.7 \times 10^{-12}$	0.047 (0.007)	$3.7 \times 10^{-12}$
<i>MLXIP1</i>	rs62466330	7	73,056,805	C/T	0.07	0.049 (0.008)	$1.2 \times 10^{-10}$	0.051 (0.007)	$5.9 \times 10^{-12}$
<i>ANK1-NKX6-3</i>	rs13266210	8	41,533,514	A/G	0.79	0.031 (0.005)	$1.3 \times 10^{-11}$	0.030 (0.004)	$1.6 \times 10^{-11}$
<i>L3MBTL3</i>	rs1415701	6	130,345,835	G/A	0.73	0.025 (0.004)	$2.6 \times 10^{-9}$	0.027 (0.004)	$4.0 \times 10^{-11}$
<i>ATAD2B</i>	rs7575873	2	23,962,647	A/G	0.88	0.038 (0.006)	$1.3 \times 10^{-11}$	0.036 (0.006)	$6.2 \times 10^{-11}$
<i>C20orf203</i>	rs28530618	20	31,275,581	A/G	0.50	0.026 (0.004)	$7.7 \times 10^{-10}$	0.024 (0.004)	$8.4 \times 10^{-10}$
<i>MAFB</i>	rs6016377	20	39,172,728	T/C	0.45	0.024 (0.004)	$9.5 \times 10^{-10}$	0.024 (0.004)	$3.7 \times 10^{-10}$
<i>CPA3</i>	rs10935733	3	148,622,968	T/C	0.42	0.022 (0.004)	$9.2 \times 10^{-9}$	0.023 (0.004)	$6.2 \times 10^{-10}$
<i>INS-IGF2</i>	rs72851023	11	2,130,620	T/C	0.07	0.048 (0.008)	$2.9 \times 10^{-10}$	0.046 (0.007)	$6.8 \times 10^{-10}$
<i>IGF2BP3</i>	rs11765649	7	23,479,013	T/C	0.76	0.027 (0.004)	$5.8 \times 10^{-10}$	0.026 (0.004)	$1.0 \times 10^{-9}$
<i>WNT4-ZBTB40</i>	rs2473248	1	22,536,643	C/T	0.87	0.033 (0.006)	$1.1 \times 10^{-8}$	0.033 (0.005)	$1.1 \times 10^{-9}$
<i>IGF1R</i>	rs7402982	15	99,193,269	A/G	0.42	0.023 (0.004)	$2.3 \times 10^{-9}$	0.023 (0.004)	$1.1 \times 10^{-9}$
<i>PLAC1</i>	rs11096402	X	133,827,868	G/A	0.25	0.028 (0.005)	$1.3 \times 10^{-9}$	N/A	N/A
<i>EBF1</i>	rs7729301	5	157,886,953	A/G	0.72	0.024 (0.004)	$1.6 \times 10^{-8}$	0.025 (0.004)	$1.3 \times 10^{-9}$
<i>SUZ12P1-CRLF3</i>	rs144843919	17	29,037,339	G/A	0.96	0.066 (0.012)	$1.4 \times 10^{-8}$	0.068 (0.011)	$1.5 \times 10^{-9}$
<i>FCGR2B</i>	rs72480273	1	161,644,871	C/A	0.17	0.031 (0.005)	$8.0 \times 10^{-9}$	0.030 (0.005)	$1.5 \times 10^{-9}$
<i>RNF219-AS1</i>	rs1819436	13	78,580,283	C/T	0.87	0.033 (0.006)	$6.3 \times 10^{-9}$	0.033 (0.005)	$1.8 \times 10^{-9}$
<i>NT5C2</i>	rs74233809	10	104,913,940	C/T	0.08	0.037 (0.007)	$5.2 \times 10^{-9}$	0.039 (0.006)	$1.8 \times 10^{-9}$
<i>SLC45A4</i>	rs12543725	8	142,247,979	G/A	0.60	0.023 (0.004)	$1.2 \times 10^{-9}$	0.022 (0.004)	$1.9 \times 10^{-9}$
<i>GPR139</i>	rs10111939	16	19,992,996	G/A	0.31	0.022 (0.004)	$1.3 \times 10^{-9}$	0.024 (0.004)	$2.7 \times 10^{-9}$
<i>SP6-SP2</i>	rs12942207	17	45,968,294	C/T	0.30	0.022 (0.004)	$5.1 \times 10^{-9}$	0.024 (0.004)	$3.0 \times 10^{-9}$
<i>GNA12</i>	rs798489	7	2,801,803	C/T	0.74	0.023 (0.004)	$2.0 \times 10^{-8}$	0.024 (0.004)	$5.0 \times 10^{-9}$
<i>PHF19</i>	rs7847628	9	123,631,225	G/A	0.67	0.023 (0.004)	$1.0 \times 10^{-8}$	0.023 (0.004)	$5.4 \times 10^{-9}$
<i>PLEKHA1</i>	rs2421016	10	124,167,512	T/C	0.48	0.021 (0.004)	$1.8 \times 10^{-8}$	0.021 (0.004)	$6.1 \times 10^{-9}$
<i>JAG1</i>	rs6040076	20	10,658,882	C/G	0.51	0.023 (0.004)	$2.0 \times 10^{-8}$	0.022 (0.004)	$7.2 \times 10^{-9}$
<i>LINC00332</i>	rs2324499	13	40,662,001	G/C	0.67	0.022 (0.004)	$7.3 \times 10^{-9}$	0.023 (0.004)	$8.3 \times 10^{-9}$
<i>IGF1</i>	rs7964361	12	102,994,878	A/G	0.08	0.039 (0.007)	$4.7 \times 10^{-9}$	0.038 (0.007)	$9.7 \times 10^{-9}$
<i>FES</i>	rs12906125	15	91,427,612	G/A	0.69	0.023 (0.004)	$1.7 \times 10^{-8}$	0.023 (0.004)	$1.0 \times 10^{-8}$
<i>TBX20</i>	rs6959887	7	35,295,365	A/G	0.61	0.023 (0.004)	$1.5 \times 10^{-8}$	0.021 (0.004)	$1.0 \times 10^{-8}$
<i>HMG2</i>	rs7742369	6	34,165,721	G/A	0.19	0.028 (0.005)	$1.0 \times 10^{-8}$	0.027 (0.005)	$1.1 \times 10^{-8}$
<i>HIST1H2BE</i>	rs9379832	6	26,186,200	A/G	0.71	0.023 (0.004)	$6.6 \times 10^{-8}$	0.024 (0.004)	$1.2 \times 10^{-8}$
<i>PTH1R</i>	rs2242116	3	46,941,116	A/G	0.39	0.022 (0.004)	$1.4 \times 10^{-8}$	0.021 (0.004)	$1.2 \times 10^{-8}$
<i>NRIP1</i>	rs2229742	21	16,339,172	G/C	0.87	0.036 (0.006)	$2.2 \times 10^{-8}$	0.034 (0.006)	$1.5 \times 10^{-8}$
<i>RB1</i>	rs2854355	13	48,882,363	G/A	0.26	0.023 (0.004)	$9.8 \times 10^{-8}$	0.024 (0.004)	$2.2 \times 10^{-8}$
<i>KREMEN1</i>	rs134594	22	29,468,456	C/T	0.35	0.023 (0.004)	$1.0 \times 10^{-8}$	0.022 (0.004)	$2.2 \times 10^{-8}$
<i>APOLD1</i>	rs11055034	12	12,890,626	C/A	0.73	0.022 (0.004)	$1.8 \times 10^{-7}$	0.023 (0.004)	$2.3 \times 10^{-8}$
<i>PEPD</i>	rs10402712	19	33,926,013	A/G	0.27	0.022 (0.004)	$4.4 \times 10^{-7}$	0.023 (0.004)	$2.3 \times 10^{-8}$
<i>ACTL9</i>	rs61154119	19	8,787,750	T/G	0.84	0.028 (0.005)	$1.1 \times 10^{-7}$	0.028 (0.005)	$2.3 \times 10^{-8}$
<i>LPAR1</i>	rs2150052	9	113,945,067	T/A	0.50	0.021 (0.004)	$2.2 \times 10^{-8}$	0.020 (0.004)	$2.8 \times 10^{-8}$
<i>ITPR2</i>	rs12823128	12	26,872,730	T/C	0.56	0.021 (0.004)	$1.9 \times 10^{-8}$	0.020 (0.004)	$3.2 \times 10^{-8}$
<i>DTL</i>	rs61830764	1	212,289,976	A/G	0.36	0.022 (0.004)	$5.6 \times 10^{-8}$	0.022 (0.004)	$4.5 \times 10^{-8}$
<i>TRIB1</i>	rs6989280	8	126,508,746	G/A	0.70	0.022 (0.004)	$2.2 \times 10^{-7}$	0.022 (0.004)	$5.0 \times 10^{-8}$
<i>MTNR1B</i>	rs10830963	11	92,708,710	G/C	0.27	0.023 (0.004)	$2.9 \times 10^{-8}$	0.022 (0.004)	$1.0 \times 10^{-7}$
<i>ABCC9</i>	rs139975827	12	22,068,161	G/A	0.63	0.025 (0.004)	$1.1 \times 10^{-8}$	0.022 (0.004)	$1.0 \times 10^{-7}$



**a.** Effects ( $\beta$  values) were aligned to the BW-raising allele. Effect allele frequency (EAF) was obtained from the trans-ancestry meta-analysis, except for *PLAC1*, for which the EAF was obtained from the European ancestry meta-analysis due to lack of X chromosome data from the non-European studies. Chr, chromosome; bp, base pair; b37, build 37; EAF, effect allele frequency; SE, standard error.

**b.** The effect of the lead SNP (absolute value of  $\beta$ , y axis) is given as a function of minor allele frequency (x axis) for 60 known (pink) and novel (green) BW loci from the trans-ancestry meta-analysis. Error bars are proportional to the standard error of the effect size. The dashed line indicates 80% power to detect association at genome-wide significance level for the sample size in trans-ancestry meta-analysis.



Extended Data Table 2 | Gene set enrichment analysis and protein–protein interaction (PPI) analysis

**a. Gene set enrichment analysis**

Database	Gene set	Number of genes (mapped to MAGENTA)	95th percentile enrichment cutoff		Expected (observed) number of genes	75th percentile enrichment cutoff		Expected (observed) number of genes
			<i>P</i>	FDR		<i>P</i>	FDR	
GOTERM	Positive regulation of glycogen biosynthetic process	10 (10)	$5.6 \times 10^{-5}$	0.005	1 (5)	$3.6 \times 10^{-3}$	0.18	3 (7)
GOTERM	Insulin-like growth factor receptor binding	13 (13)	$2.4 \times 10^{-5}$	0.006	1 (6)	0.02	0.35	3 (7)
GOTERM	Positive regulation of glucose import	22 (22)	$1.0 \times 10^{-4}$	0.019	1 (7)	0.02	0.36	6 (10)
GOTERM	Insulin receptor signalling pathway	35 (34)	$2.8 \times 10^{-5}$	0.022	2 (9)	$4.3 \times 10^{-3}$	0.27	9 (16)
GOTERM	Chromatin remodelling complex	11 (9)	$9.0 \times 10^{-4}$	0.036	0 (4)	0.16	0.55	2 (4)
KEGG	Glycosphingolipid biosynthesis globo-series	14 (13)	$2.6 \times 10^{-3}$	0.037	1 (4)	0.21	0.48	3 (5)
KEGG	Melanoma	71 (67)	$1.6 \times 10^{-3}$	0.037	3 (10)	0.05	0.35	17 (23)
KEGG	Terpenoid backbone biosynthesis	15 (15)	$5.9 \times 10^{-3}$	0.039	1 (1)	0.15	0.44	4 (6)
KEGG	Type 2 Diabetes Mellitus	47 (45)	$2.2 \times 10^{-3}$	0.040	2 (8)	0.14	0.46	11 (15)
Panther	Cholesterol biosynthesis	11 (11)	$1.8 \times 10^{-3}$	0.040	1 (4)	0.29	0.64	3 (4)
BIOCARTA	Growth hormone pathway	28 (27)	$3.0 \times 10^{-4}$	0.044	1 (7)	0.11	0.25	7 (10)
KEGG	Oocyte meiosis	114 (108)	$1.0 \times 10^{-3}$	0.048	5 (14)	0.07	0.45	27 (34)
<b>Custom gene set of imprinted genes</b>								
GTEX	Imprinted genes (All)	77 (72)	$1.9 \times 10^{-4}$	-	4 (12)	0.11	-	18 (23)
GTEX	Imprinted genes (Primary)	38 (35)	$6.9 \times 10^{-3}$	-	2 (6)	0.14	-	9 (12)
GTEX	Imprinted genes (Primary + Suggestive)	55 (50)	0.010	-	3 (7)	0.25	-	13 (15)

**b. Protein-protein interaction analysis**

		Number of genes (overlapped with PPI network)			
Database	Pathway		Z score	P	adjusted P <sup>a</sup>
GOTERM	Epidermal growth factor receptor signalling pathway	198 (31)	7.97	3.3x10 <sup>-10</sup>	1.4x10 <sup>-7</sup>
GOTERM	Insulin receptor signalling pathway	151 (26)	7.90	1.1x10 <sup>-9</sup>	2.9x10 <sup>-7</sup>
GOTERM	Stimulatory C-type lectin receptor signalling pathway	121 (22)	7.59	7.5x10 <sup>-9</sup>	1.2x10 <sup>-6</sup>
GOTERM	Negative regulation of canonical Wnt signalling pathway	152 (25)	7.46	6.2x10 <sup>-9</sup>	1.2x10 <sup>-6</sup>
GOTERM	Notch signalling pathway	129 (22)	7.21	2.6x10 <sup>-8</sup>	3.3x10 <sup>-6</sup>
GOTERM	Cellular response to insulin stimulus	71 (16)	7.62	3.7x10 <sup>-8</sup>	4.1x10 <sup>-6</sup>
GOTERM	Positive regulation of glycogen biosynthetic process	15 (8)	9.39	5.3x10 <sup>-8</sup>	5.1x10 <sup>-6</sup>
GOTERM	Positive regulation of protein phosphorylation	114 (20)	7.03	6.8x10 <sup>-8</sup>	5.9x10 <sup>-6</sup>
GOTERM	Positive regulation of glucose import	27 (10)	8.42	8.3x10 <sup>-8</sup>	6.5x10 <sup>-6</sup>
GOTERM	Fc-epsilon receptor signalling pathway	186 (26)	6.58	9.6x10 <sup>-8</sup>	6.8x10 <sup>-6</sup>

Two complementary analyses of the overall GWAS summary data identified enrichment of BW associations in biological pathways related to metabolism, growth and development. **a**, The top results (FDR < 0.05 at the 95th percentile enrichment threshold) from a total of 3,216 biological pathways tested for enrichment of multiple modest associations with BW. Additionally, results are shown for custom sets of imprinted genes: Primary, genes identified as highly likely to be imprinted in the GTEx database (tested  $n = 38$ ); Primary + suggestive, genes identified as highly likely and suggestively imprinted in GTEx ( $n = 55$ ); All, the above plus genes selected from the literature where imprinting status is consistent in GTEx ( $n = 77$ ). **b**, The results of a complementary analysis of empirical PPI data, displaying the top 10 most significant pathways enriched for BW-association scores.

<sup>a</sup>*P* value is adjusted for multiple correction using the Benjamini–Hochberg method.

# A cholinergic basal forebrain feeding circuit modulates appetite suppression

Alexander M. Herman<sup>1</sup>, Joshua Ortiz-Guzman<sup>1</sup>, Mikhail Kochukov<sup>2</sup>, Isabella Herman<sup>1,3</sup>, Kathleen B. Quast<sup>2</sup>, Jay M. Patel<sup>3,4</sup>, Burak Tepe<sup>1</sup>, Jeffrey C. Carlson<sup>1</sup>, Kevin Ung<sup>1</sup>, Jennifer Selever<sup>2,5</sup>, Qingchun Tong<sup>6</sup> & Benjamin R. Arenkiel<sup>1,2,4,5</sup>

**Atypical food intake is a primary cause of obesity and other eating and metabolic disorders. Insight into the neural control of feeding has previously focused mainly on signalling mechanisms associated with the hypothalamus<sup>1–5</sup>, the major centre in the brain that regulates body weight homeostasis<sup>6,7</sup>. However, roles of non-canonical central nervous system signalling mechanisms in regulating feeding behaviour have been largely uncharacterized. Acetylcholine has long been proposed to influence feeding<sup>8–10</sup> owing in part to the functional similarity between acetylcholine and nicotine, a known appetite suppressant. Nicotine is an exogenous agonist for acetylcholine receptors, suggesting that endogenous cholinergic signalling may play a part in normal physiological regulation of feeding. However, it remains unclear how cholinergic neurons in the brain regulate food intake. Here we report that cholinergic neurons of the mouse basal forebrain potentially influence food intake and body weight. Impairment of cholinergic signalling increases food intake and results in severe obesity, whereas enhanced cholinergic signalling decreases food consumption. We found that cholinergic circuits modulate appetite suppression on downstream targets in the hypothalamus. Together our data reveal the cholinergic basal forebrain as a major modulatory centre underlying feeding behaviour.**

The diagonal band of Broca (DBB) (Extended Data Fig. 1a–g) is a major component of the cholinergic basal forebrain<sup>11</sup>. Studies have shown that cell types associated with feeding are connected with the DBB<sup>12,13</sup>. Given this correlation, we sought to determine if cholinergic neurons in the mouse DBB modulate feeding. With feeding after an overnight fast, we observed co-localization between c-Fos and the cholinergic-specific marker, choline acetyltransferase (ChAT) (Fig. 1a–f). Given their activation in response to food intake, we next assessed the necessity of cholinergic DBB neurons in feeding modulation. We stereotactically targeted cholinergic DBB neurons for ablation using a Cre-dependent<sup>14</sup> adeno-associated virus (AAV) that expressed an enhanced yellow fluorescent protein (EYFP)-linked diphtheria toxin receptor in *Chat-cre<sup>+/+</sup>; R26<sup>LSL-tdTomato/+</sup>* mice. DBB cholinergic neurons were mainly decimated after treatment with diphtheria toxin, whereas those in surrounding areas of the brain were unaffected (Fig. 1g–n). Within two weeks after ablation, we observed hyperphagia (Fig. 1o) and severe obesity (Fig. 1p–q). To assess the extent that food consumption contributed to this phenotype, we conducted paired feeding assays on obese DBB-ablated animals versus their non-ablated controls. Control mice exhibited normal food intake and minimal weight gain, whereas DBB-ablated animals lost weight when available food was limited to control levels (Fig. 1r–s), suggesting that hyperphagia was required to maintain obesity. This effect did not appear to be mediated by changes in energy expenditure, as overall levels of activity (Fig. 1t) and oxygen consumption (Extended Data Fig. 2a) were unaffected during early stages post-DBB ablation. Metabolic

activity was only altered once obesity became a burden (Extended Data Fig. 2b–k).

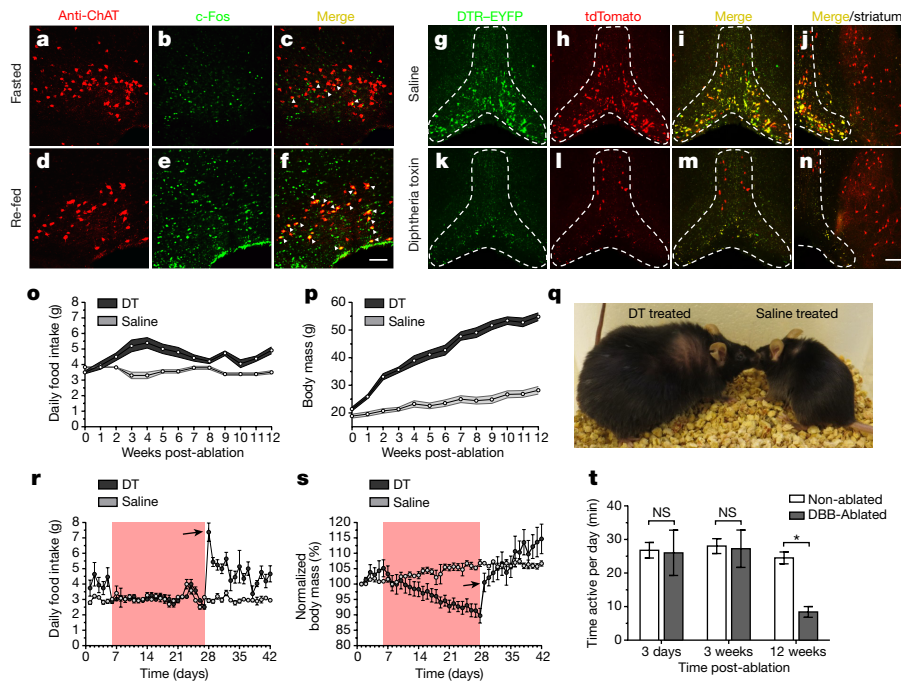
We sought to determine whether conditional removal of cholinergic neurotransmission selectively from the DBB, without cell death, led to increased food consumption and obesity. We stereotactically delivered an EGFP-Cre-expressing AAV into the DBB of *Chat<sup>loxP/loxP</sup>* homozygous mice (Fig. 2a), and evaluated ChAT knockout by immunohistochemistry (Fig. 2b–k). Counts of ChAT-immunopositive neurons from control and experimental brains showed a 72% and 55% decrease in ChAT expression in the DBB of Cre-expressing female and male mice, respectively, compared to controls (Fig. 2l). Consistent with a role for cholinergic signalling in feeding behaviour, Cre-expressing animals displayed increased food intake (Fig. 2m) and subsequent weight gain (Fig. 2n). Because knockout is restricted to the *Chat* gene, these data imply that the observed effects on feeding and body weight were mediated by cholinergic signalling, although it does not rule out a role for GABAergic neurotransmission from cholinergic DBB neurons, which have recently been reported to co-release GABA<sup>15</sup> (Extended Data Fig. 3a–c).

To test how acute inactivation of these neurons affected food intake, *Chat-cre<sup>+/+</sup>* mice were DBB-targeted for conditional expression of hM4D-EGFP (Extended Data Fig. 4a). Following CNO treatment, hM4D-EGFP-expressing mice consumed more food compared to controls (Extended Data Fig. 4b), suggesting that both acute and chronic inactivation of cholinergic DBB neurons were sufficient to increase food intake.

We sought to determine if activating DBB cholinergic neurons was sufficient to suppress food intake *in vivo*. To test this, we expressed channelrhodopsin-2 (ChR2) in DBB cholinergic neurons, followed by monitoring feeding behaviour during photostimulation (Fig. 3a–g). We found that prolonged (48 h) activation of cholinergic DBB neurons significantly decreased daily food intake under conditions of normal food access (Fig. 3h, i), whereas mock-stimulated, non-ChR2-expressing controls showed no significant changes in feeding behaviour (average daily food intake was 4.26 g ( $\pm 0.22$  g, s.e.m.); 112.48% of baseline ( $\pm 5.91\%$ , s.e.m.),  $n = 3$ ). To test the acute effect of DBB stimulation on feeding, we fasted ChR2-expressing mice overnight and presented them with chow in the morning. We observed that acute cell body stimulation resulted in a similar overall decrease (approximately 25%) in food intake (Fig. 3j–k).

We selectively targeted cholinergic DBB neurons for conditional viral expression of a presynaptically localized, synaptophysin-EGFP (Extended Data Fig. 5a–e), allowing the identification of their presumptive downstream synaptic targets. Using this method, we observed DBB neuron terminals in previously reported areas of the brain<sup>11</sup> (Extended Data Fig. 5f–n), as well as in ventral-medial domains of the hypothalamus (Extended Data Fig. 5o–u), which have previously been shown to express markers for cholinergic

<sup>1</sup>Program in Developmental Biology, Baylor College of Medicine, Houston, Texas 77030, USA. <sup>2</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. <sup>3</sup>Medical Scientist Training Program, Baylor College of Medicine, Houston, Texas 77030, USA. <sup>4</sup>Department of Neuroscience, Baylor College of Medicine, Houston, Texas 77030, USA. <sup>5</sup>Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Houston, Texas 77030, USA. <sup>6</sup>Institute of Molecular Medicine, University of Texas Health Science Center, Houston, Texas 77030, USA.



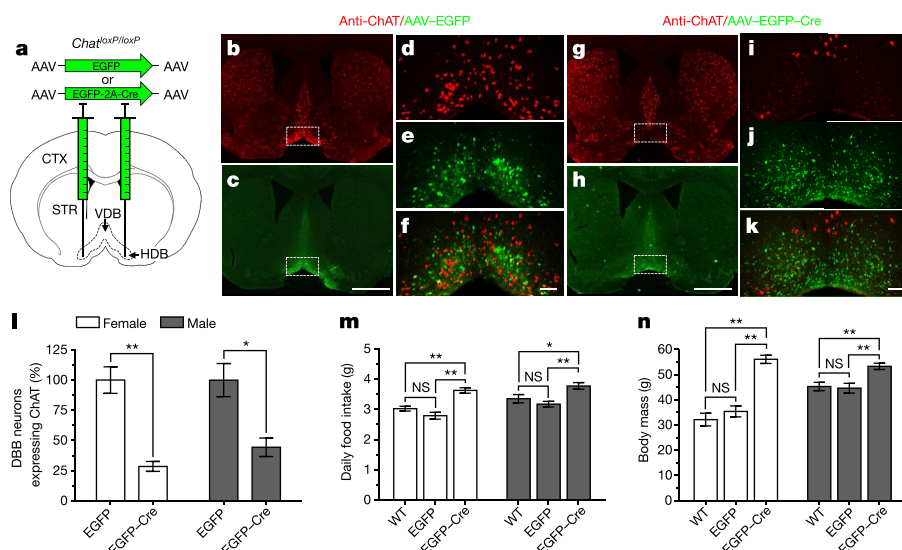
**Figure 1 | DBB-specific cholinergic cell death results in hyperphagia and obesity.** **a–f**, c-Fos and ChAT immunohistochemistry in the HDB. **g–n**, DBB of saline-treated (**g–j**) or diphtheria toxin treated (**k–n**) animals. Scale bars, 100  $\mu$ m. **o**, Food intake between DBB-ablated ( $n = 15$  mice) and non-ablated animals ( $n = 13$  mice). Time points represented as mean  $\pm$  s.e.m.,  $P = 0.0002$ ,  $F(1, 26) = 18.06$  by two-way ANOVA with repeated measures. **p**, Body weight between DBB-ablated ( $n = 15$ ) and non-ablated animals ( $n = 13$ ). Time points represented as mean  $\pm$  s.e.m.,

$P < 0.0001$ ,  $F(1, 26) = 143.1$  by two-way ANOVA with repeated measures. **q**, Representative DBB-ablated and non-ablated mice. **r**, **s**, Average daily food intake (**r**) and body mass (normalized with respect to the day-0 mass; **s**) during paired feeding ( $n = 5$  mice per group), pink shading represents restrictive period. **t**, Total time active per day before obesity, and during early and late stages of obesity. Data represented as mean  $\pm$  s.e.m.,  $*P < 0.05$  by two-sided, unpaired Student's  $t$ -test.

terminals of unknown origin<sup>16</sup>. Within this region resides the arcuate nucleus of the hypothalamus, a site that contains well-characterized cell populations that conversely regulate hunger and satiety through NPY/AgRP-expressing and POMC-expressing neurons, respectively<sup>17</sup>. As cholinergic basal forebrain neurons co-express GABA, it is possible that loss of cholinergic neurons from the DBB could disinhibit AgRP neurons, thereby promoting food intake through elevated expression levels of *Agrp*. To test this, *Agrp* transcripts were assessed in DBB-ablated animals that were pair-fed to respective controls shortly after ablation, before hyperphagic-induced obesity. Under these conditions, *Agrp* transcript analysis from the arcuate nucleus of DBB-ablated animals showed no significant decrease in levels of *Agrp* compared to controls (Extended Data Fig. 6a). Interestingly, however, *Pomc* transcript

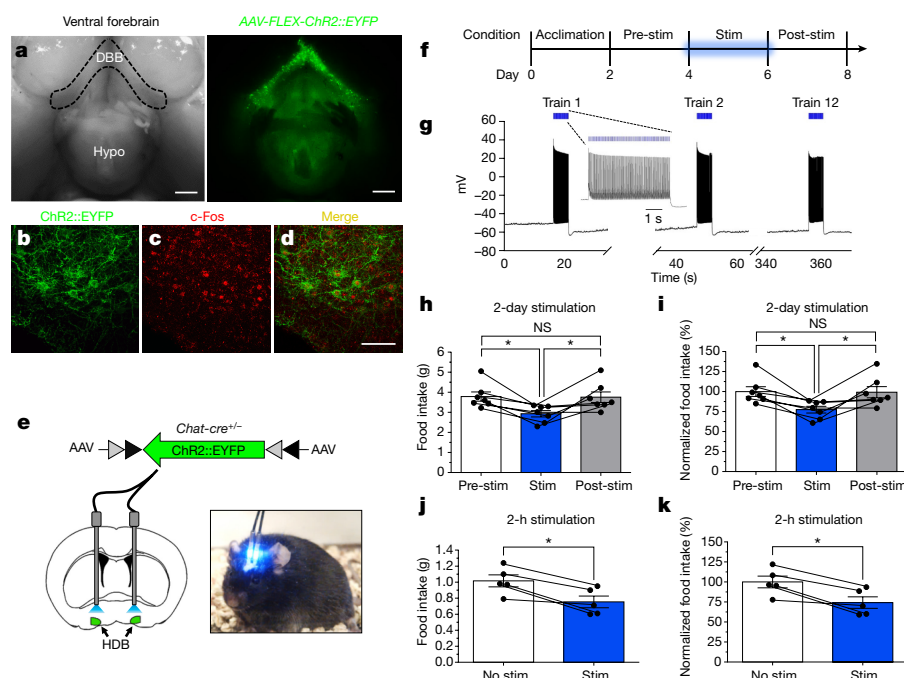
levels were strikingly reduced in DBB-ablated mice (Extended Data Fig. 6a), suggesting that cholinergic DBB neurons may normally act to modulate downstream satiety pathways. Previous reports show that POMC neurons are activated by cholinergic nicotinic agonists, however, their effect on AgRP/NPY neurons appears much more variable. From our own recordings (Extended Data Fig. 6b), acetylcholine did not show fast excitation on NPY neurons, perhaps due to muscarinic acetylcholine receptor (mAChR) activity or through broader network effects.

Recent studies have implicated the arcuate nucleus in appetite suppression via AChR signalling onto POMC neurons<sup>18</sup>, and conditional rabies virus tracing experiments suggested that arcuate POMC neurons receive significant input from neurons in the DBB<sup>12</sup>. Whole-cell



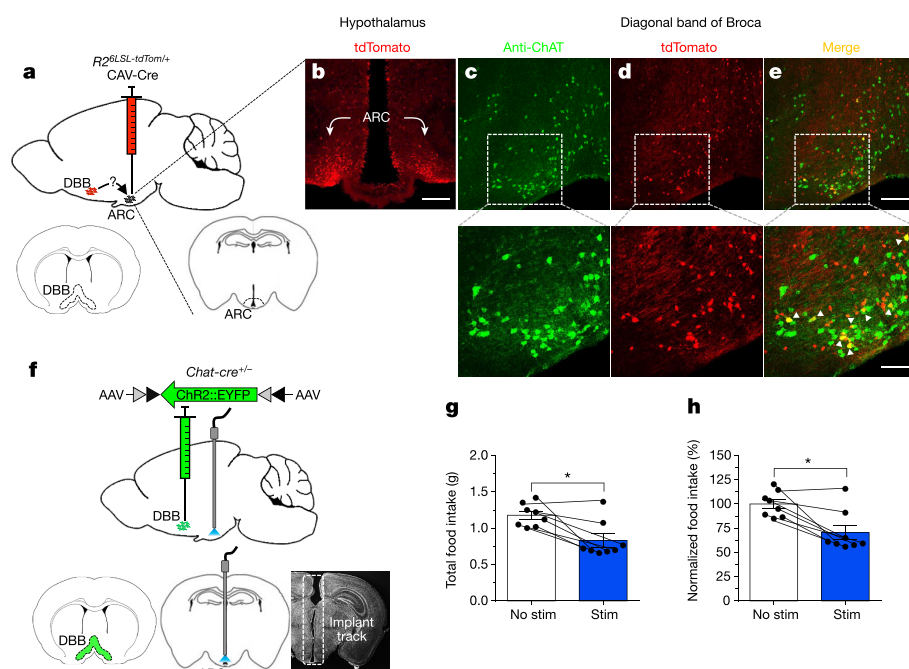
**Figure 2 | Chat conditional knockout from the DBB leads to obesity and hyperphagia.** **a**, Strategy for *Chat* conditional knockout. **b–k**, ChAT immunostains of DBB in EGFP-expressing (**b–f**) or EGFP-Cre-expressing (**g–k**) *Chat<sup>loxP/loxP</sup>* brains. Scale bars represent 1 mm (**c**, **h**) or 100  $\mu$ m (**f**, **k**). **l**, Percentage of DBB neurons expressing ChAT ( $n = 3$  per gender). Data represented as mean  $\pm$  s.e.m.,  $*P < 0.05$ ,  $**P < 0.01$  by two-sided, unpaired Student's  $t$ -test. **m**, **n**, Daily average food intake (**m**) and average body mass (**n**) (12 weeks after the conditional knockout,  $n = 7$  females, 5 males for EGFP-Cre or EGFP mice,  $n = 6$  females, 6 males for wild type). Data represented as mean  $\pm$  s.e.m.,  $*P < 0.05$ ,  $**P < 0.01$  by two-sided, unpaired Student's  $t$ -test. Schematic in **a** was adapted from an image from the Allen Brain Institute Reference Atlas<sup>30</sup>.





recordings from labelled POMC neurons in *Pomc-EGFP*<sup>+/−</sup> mice showed that neuronal firing increased in the presence of acetylcholine (Extended Data Fig. 6c–d), and this response was blocked by AChR blockers (Extended Data Fig. 6e). To assess potential connectivity between the DBB and arcuate nucleus, we implemented retrogradely transported canine adenovirus (CAV-2)<sup>19</sup>. Using conditional *R26<sup>LSL-tdTomato</sup>*<sup>+</sup> mice, we delivered CAV-Cre into the arcuate nucleus, and found that cholinergic DBB neurons provide input (Fig. 4a–e). Non-cholinergic cell types in the DBB were also labelled, suggesting that both cholinergic and non-cholinergic cell types from the DBB provide input into the hypothalamus to regulate its functions. To investigate this connectivity *in vivo*, *Chat-cre*<sup>+/−</sup> mice were targeted for conditional ChR2::EYFP expression in DBB cholinergic neurons, and a fibre optic was implanted into the ventral space of the third ventricle to stimulate DBB terminals within the ventral

hypothalamus (Fig. 4f). ChR2-expressing and non-ChR2-expressing (mock-stimulated) implanted animals were fasted overnight and presented with chow for two hours. Mock-stimulated animals showed no significant changes in feeding behaviour during photostimulation (average was 1.381 g ( $\pm 0.1429$  g, s.e.m.); 117.1% of baseline ( $\pm 12.12\%$ , s.e.m.),  $n = 6$  animals), whereas stimulated ChR2-expressing animals exhibited diminished food intake compared to non-stimulated conditions (Fig. 4g–h), consistent with decreases observed from DBB cell body stimulation, and supporting anatomical tracing data. Although these data support a role for cholinergic modulation of the arcuate nucleus, given sparse labelling of terminals throughout the hypothalamus, as well as dense innervation of the median eminence (Extended Data Fig. 7a–h), these results do not rule out an influence of other hypothalamic sites and cell types. Lastly, to test if this effect was AChR-mediated, we systemically injected





photostimulated animals with the nicotinic cholinergic receptor (nAChR) antagonist mecamylamine, which has previously been shown to block nAChR-mediated decreases in food intake<sup>18</sup>. In the presence of mecamylamine, photostimulated animals showed a blunted response compared to stimulation alone (Extended Data Fig. 8a–b), suggesting that decreased food intake after stimulation was mediated in part by AChR signalling. Notably, however, we did not observe a complete phenotypic blockade, suggesting that mAChR signalling may play a synergistic role in modulating food intake, and may therefore explain why nAChR antagonism alone was not sufficient to fully suppress feeding.

Together, our data demonstrate a powerful role for cholinergic basal forebrain neurons in modulating food intake. Cholinergic systems are highly druggable, which may offer a new avenue for targeting cholinergic mechanisms to help treat eating disorders. AChR transcript analysis from the arcuate nucleus showed expression of common AChR subunits (Extended Data Fig. 9a), serving as possible targets for pharmacological interventions. Understanding the AChR profiles of prominent feeding-associated cell types will surely be useful in this respect<sup>20</sup>. Additionally, revealing the contribution of other signalling mechanisms from cholinergic neurons will be essential in understanding their composite roles in feeding. Cholinergic and other neuromodulatory systems are known to co-express multiple neurotransmitters and/or neuropeptides<sup>15,21,22</sup>. Although cholinergic signalling is sure to play an important role, it remains unclear how dual transmitter systems function in concert to modulate their targets. Although our studies indicate a modulatory effect of acetylcholine on POMC neurons to influence food intake, it is possible that other cell types of the hypothalamus and/or other brain regions may be affected in a manner yet to be determined by cholinergic innervation. As such, it is possible that our observed phenotypes are not due solely to downstream effects in the arcuate nucleus alone, but may be a combinatorial effect of altered cholinergic signalling at other projection sites or cell types of the hypothalamus or other innervated brain regions. It will be especially interesting to determine the role of alternative cholinergic centres in modulating feeding behaviour such as those associated with the reward and addiction pathways of the mesolimbic system. The boundary between homeostatic and non-homeostatic (hedonic/aversive) control of feeding is not always distinct or clear-cut, and often converges<sup>23,24</sup>. Given its role as a powerful modulator of reward<sup>25–28</sup> and aversion<sup>26,27,29</sup>, and its involvement in addiction<sup>26,27</sup>, acetylcholine is a prime candidate that may link known canonical homeostatic mechanisms that govern feeding behaviour, with the reward and/or aversive aspects of food intake. Nonetheless, fully understanding the relationship between cholinergic systems and their role in modulating feeding behaviour is an important step towards uncovering a comprehensive profile of feeding regulation.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 31 March; accepted 19 August 2016.**

**Published online 3 October 2016.**

- Aponte, Y., Atasoy, D. & Sternson, S. M. AGRP neurons are sufficient to orchestrate feeding behavior rapidly and without training. *Nat. Neurosci.* **14**, 351–355 (2011).
- Chen, Y., Lin, Y. C., Kuo, T. W. & Knight, Z. A. Sensory detection of food rapidly modulates arcuate feeding circuits. *Cell* **160**, 829–841 (2015).
- Luquet, S., Perez, F. A., Hnasko, T. S. & Palmiter, R. D. NPY/AgRP neurons are essential for feeding in adult mice but can be ablated in neonates. *Science* **310**, 683–685 (2005).
- Pinto, S. *et al.* Rapid rewiring of arcuate nucleus feeding circuits by leptin. *Science* **304**, 110–115 (2004).
- Garfield, A. S. *et al.* A neural basis for melanocortin-4 receptor-regulated appetite. *Nat. Neurosci.* **18**, 863–871 (2015).
- Morton, G. J., Cummings, D. E., Baskin, D. G., Barsh, G. S. & Schwartz, M. W. Central nervous system control of food intake and body weight. *Nature* **443**, 289–295 (2006).
- Elmquist, J. K., Elias, C. F. & Saper, C. B. From lesions to leptin: hypothalamic control of food intake and body weight. *Neuron* **22**, 221–232 (1999).
- Pistelli, F., Aquilini, F. & Carrozzi, L. Weight gain after smoking cessation. *Monaldi Arch. Chest Dis.* **71**, 81–87 (2009).
- Fulkerson, J. A. & French, S. A. Cigarette smoking for weight loss or control among adolescents: gender and racial/ethnic differences. *J. Adolesc. Health* **32**, 306–313 (2003).
- Voorhees, C. C., Schreiber, G. B., Schumann, B. C., Biro, F. & Crawford, P. B. Early predictors of daily smoking in young women: The National Heart, Lung, and Blood Institute growth and health study. *Prev. Med.* **34**, 616–624 (2002).
- Zaborszky, L., van den Pol, A. & Gyengesi, E. In *The Mouse Nervous System* (eds Watson C., Paxinos, G. & Puelles, L.) 684–718 (2012).
- Wang, D. *et al.* Whole-brain mapping of the direct inputs and axonal projections of POMC and AgRP neurons. *Front. Neuroanat.* **9**, 40 (2015).
- Sakurai, T. *et al.* Input of orexin/hypocretin neurons revealed by a genetically encoded tracer in mice. *Neuron* **46**, 297–308 (2005).
- Atasoy, D., Aponte, Y., Su, H. H. & Sternson, S. M. A FLEX switch targets Channelrhodopsin-2 to multiple cell types for imaging and long-range circuit mapping. *J. Neurosci.* **28**, 7025–7030 (2008).
- Saunders, A., Granger, A. J. & Sabatini, B. L. Corelease of acetylcholine and GABA from cholinergic forebrain neurons. *eLife* **4**, e06412 (2015).
- Ichikawa, T., Ajiki, K., Matsuura, J. & Misawa, H. Localization of two cholinergic markers, choline acetyltransferase and vesicular acetylcholine transporter in the central nervous system of the rat: in situ hybridization histochemistry and immunohistochemistry. *J. Chem. Neuroanat.* **13**, 23–39 (1997).
- Gropp, E. *et al.* Agouti-related peptide-expressing neurons are mandatory for feeding. *Nat. Neurosci.* **8**, 1289–1291 (2005).
- Mineur, Y. S. *et al.* Nicotine decreases food intake through activation of POMC neurons. *Science* **332**, 1330–1332 (2011).
- Soudais, C., Laplace-Builhe, C., Kissa, K. & Kremer, E. J. Preferential transduction of neurons by canine adenovirus vectors and their efficient retrograde transport *in vivo*. *FASEB J.* **15**, 2283–2285 (2001).
- Henry, F. E., Sugino, K., Tozer, A., Branco, T. & Sternson, S. M. Cell type-specific transcriptomics of hypothalamic energy-sensing neuron responses to weight-loss. *eLife* **4**, e09800 (2015).
- Ren, J. *et al.* Habenula “cholinergic” neurons co-release glutamate and acetylcholine and activate postsynaptic neurons via distinct transmission modes. *Neuron* **69**, 445–452 (2011).
- Saunders, A. *et al.* A direct GABAergic output from the basal ganglia to frontal cortex. *Nature* **521**, 85–89 (2015).
- Saper, C. B., Chou, T. C. & Elmquist, J. K. The need to feed: homeostatic and hedonic control of eating. *Neuron* **36**, 199–211 (2002).
- Liu, C., Lee, S. & Elmquist, J. K. Circuits controlling energy balance and mood: inherently intertwined or just complicated intersections? *Cell Metab.* **19**, 902–909 (2014).
- Wise, R. A. Brain reward circuitry: insights from unsensed incentives. *Neuron* **36**, 229–240 (2002).
- Picciotto, M. R., Higley, M. J. & Mineur, Y. S. Acetylcholine as a neuromodulator: cholinergic signaling shapes nervous system function and behavior. *Neuron* **76**, 116–129 (2012).
- Miwa, J. M., Freedman, R. & Lester, H. A. Neural systems governed by nicotinic acetylcholine receptors: emerging hypotheses. *Neuron* **70**, 20–33 (2011).
- Hangya, B., Ranade, S. P., Lorenc, M. & Kepecs, A. Central cholinergic neurons are rapidly recruited by reinforcement feedback. *Cell* **162**, 1155–1168 (2015).
- Hoebel, B. G., Avena, N. M. & Rada, P. Accumbens dopamine-acetylcholine balance in approach and avoidance. *Curr. Opin. Pharmacol.* **7**, 617–627 (2007).
- Website: © 2015 Allen Institute for Brain Science. Allen Mouse Brain Atlas [Internet]. Available from: <http://mouse.brain-map.org>.

**Acknowledgements** This study was supported by NIH grants 5F31NS089411 to A.M.H., R01NS078294 to B.R.A., R01DK109934 to B.R.A. and Q.T., P30DK079638 to the BCM Mouse Metabolic Core, and U54HD083092 to the BCM IDRC. Support was also provided to B.R.A. from the Klarman Family Foundation, the Klingenstein-Simons Fellowship Award, the Brain and Behavior Research Foundation, the Charif Souki Fund, and the McNair Medical Institute. We thank H. Zoghbi, H. Bellen, M. Wang, and M. Krashes for input on this manuscript.

**Author Contributions** B.R.A., A.M.H., and Q.T. designed all experiments. J.C.C., J.O.-G., A.M.H., I.H., M.K., J.M.P., K.Q., B.T., and K.U. performed experiments. B.R.A., A.M.H., and K.Q. analysed the data. Genetic reagents and viral constructs were engineered by B.R.A., A.M.H., I.H., and J.S. A.M.H. and B.R.A. wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.R.A. ([arenkiel@bcm.edu](mailto:arenkiel@bcm.edu)).

**Reviewer Information** *Nature* thanks L. de Lecea and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

**Data reporting.** No statistical methods were used to predetermine sample size. When relevant (such as experiments requiring multiple trials), randomization was carried out. Behavioural trial experiments were randomized. There was blinding of initial allocation of animals into groups, but not thereafter. Cell counts were blinded.

**Experimental mouse lines.** All animals used in this study were treated in compliance with US Department of Health and Human Services and Baylor College of Medicine IACUC guidelines. For the studies reported here, both male and female mice were considered for analyses. Standard pellet mouse chow (Harlan, 2920X) was used for all experiments, and all animals were maintained on a normal 12-h light–dark cycle. *Chat-cre* (B6;129S6-*Chat<sup>tm2(cre)Lowl/J</sup>*), *Pomc-EGFP* (C57BL/6J-Tg(Pomc-EGFP)1Low/J), *Npy-hrGFP* (B6.FVB-Tg(Npy-hrGFP)1Lowl/J), *Chat<sup>loxP/loxP</sup>* (B6.129-*Chat<sup>tm1Jrs/J</sup>*), *R26<sup>LSL-tdTomato</sup>* (B6.Cg-Gt(ROSA)26Sor<sup>tm14(CAG-tdTomato)Hze/J</sup>), and *Vgat-cre* (Slc32a1<sup>tm2(cre)Lowl/J</sup>) mice were originally purchased and are available from Jackson Laboratories. *Chat-cre<sup>+/-</sup>*; *R26<sup>LSL-tdTomato/+</sup>* mice were generated by crossing heterozygous male *Chat-cre<sup>+/-</sup>* mice with female homozygous *R26<sup>LSL-tdTomato</sup>* mice. *Chat<sup>loxP/loxP</sup>* animals were bred and maintained as homozygotes. A *Vgat-cre* homozygous male was crossed to female C57BL/6J mice to generate heterozygous *Vgat-cre<sup>+/-</sup>* animals. *Chat-cre<sup>+/-</sup>*, *Pomc-EGFP<sup>+/-</sup>* and *Npy-hrGFP<sup>+/-</sup>* mice used in this study were maintained as heterozygotes and bred to wild-type C57BL/6J female mice. Genotyping for *Chat<sup>loxP/loxP</sup>*, *Pomc-EGFP*, and *Npy-hrGFP* animals was done according to available Jackson Laboratory protocols for these strains. Genotyping for Cre was done using primers for Cre recombinase detection (forward primer: 5'-GCATTCTGGGGATTGCTTA-3', reverse primer: 5'-GTCATCCTTAGCGCCGTA-3').

**Microscopy and immunohistochemistry.** Animals were deeply anaesthetized using isoflurane and were transcardially perfused with PBS followed by 10% neutral buffered formalin (NBF, Azar Scientific). Brains were dissected and post-fixed in 10% NBF overnight at 4°C. Brains were cryoprotected in a 20% sucrose/PBS solution at 4°C for one day, followed by a 30% sucrose/PBS solution at 4°C for one more day. Brains were then embedded and frozen in OCT and stored at -80°C. Brains visualized using endogenous or virally-expressing fluorescent reporters were cut using a cryostat (Leica CM1860) in coronal sections at 25–30 µm. For ChAT and β-endorphin immunohistochemistry, 40 µm free-floating sections were blocked for 1 h at room temperature in 10% horse serum blocking solution, made in PBS-TC (1 × PBS, 0.5% Triton-X 100, 0.1 mM CaCl<sub>2</sub>, pH 7.35). Sections were then incubated overnight at 4°C at a 1:200 dilution of block solution containing goat anti-ChAT primary antibody (Millipore, AB144P) or rabbit anti-β-endorphin primary antibody (Phoenix Pharmaceuticals, H-022-33). Sections were then washed 4 times, 30 min each in plain PBS-TC. Sections were then incubated in secondary antibody (donkey anti-goat Alexafluor-488 or Alexafluor-555, Life Technologies) at a 1:200 dilution for 3 h at room temperature. Sections were then washed 4 times for 30 min each in PBS-TC. All sections were mounted using DAPI Fluoromount-G (Southern Biotech, 0100-20). Detection of fluorescent expression was performed using a Leica TCS SPE confocal microscope under a 10× or 20× objective.

**Stereotaxic injections and viral constructs.** For all stereotaxic injections, mice were anesthetized using a ketamine/dormitor mixture and were maintained under anaesthesia using vaporized isoflurane with O<sub>2</sub>. All injections were performed using a stereotaxic apparatus synced to Angle Two software for coordinate guidance. For DTR-mediated cell death of cholinergic neurons, female *Chat-cre<sup>+/-</sup>* or *Chat-cre<sup>+/-</sup>*; *R26<sup>LSL-tdTomato/+</sup>* mice (8–10 weeks old) were bilaterally injected into the horizontal limb of the diagonal band of Broca (HDB, right hemisphere, from bregma: AP = +0.14, DV = -5.80, ML = -1.29; left hemisphere, from bregma: AP = +0.14, DV = -5.74, ML = +1.17) with 500 nl per hemisphere of a Cre-dependent AAV-EF1α-FLEX-DTR-P2A-EYFP-WPRE-hGHPa, serotype DJ/8. For conditional *Chat* knockout experiments, *Chat<sup>loxP/loxP</sup>* animals (8–10 weeks old) were bilaterally injected into the HDB with 300 nl per hemisphere of AAV-EF1α-EGFP-P2A-CRE-WPRE-hGHPa for experimental animals, or AAV-EF1α-EGFP-WPRE-hGHPa for control animals (serotype DJ/8 for both AAVs). For synaptophysin tracing experiments using the EGFP variant, *Chat-cre<sup>+/-</sup>* mice (8–16 weeks old) were injected bilaterally into the HDB with 500 nL per hemisphere of a Cre-dependent AAV-EF1α-FLEX-Syn::EGFP-WPRE-hGHPa, serotype DJ/8. For synaptophysin tracing experiments using the mRuby2 variant, *Chat-cre<sup>+/-</sup>*; *Pomc-EGFP<sup>+/-</sup>* mice (12 weeks old) were injected bilaterally into the HDB with 500 nl per hemisphere of a Cre-dependent AAV-EF1α-FLEX-Syn::mRuby2-WPRE-hGHPa, serotype DJ/8. Lastly, for *in vivo* ChR2 behaviour experiments, male *Chat-cre<sup>+/-</sup>* mice (12–14 weeks old) were bilaterally injected into the HDB with 500 nl per hemisphere of a Cre-dependent AAV-EF1α-DIO-hChR2(H134R)-EYFP-WPRE-hGHPa (Addgene, plasmid number 20298) serotype 2/9.

**Neuronal activation after feeding and c-Fos IHC.** Male wild-type animals were fasted overnight before being presented with standard pellet mouse chow (Harlan, 2920X) for 3 h the following morning (fed group), while a second group of mice was

fasted overnight but not presented with chow (fasted group). Mice were then immediately euthanized and perfused with PBS and 10% NBF. Brains were fixed overnight in 10% NBF before two overnight fixations in 20% and 30% sucrose/PBS solutions. Brains were frozen in OCT cutting compound and cryosectioned at 35 µm. Sections were then blocked for 1 h at room temperature in 10% horse serum blocking solution, made in PBS-TC (1 × PBS, 0.5% Triton-X 100, 0.1 mM CaCl<sub>2</sub>, pH 7.35). Sections were then incubated overnight at 4°C at a 1:200 dilution of block solution containing goat anti-ChAT primary antibody (Millipore, AB144P) and 1:500 dilution of rabbit anti-c-Fos antibody (Calbiochem, PC38). Sections were then washed 4 times for 30 min each in plain PBS-TC. Sections were then incubated in secondary antibodies (donkey anti-goat Alexafluor-488, and donkey anti-rabbit Alexafluor-546) at a 1:200 dilution each for 3 h at room temperature. Sections were then washed 4 times for 30 min each in PBS-TC. All sections were mounted using DAPI Fluoromount-G (Southern Biotech, 0100-20). Detection of fluorescent expression was performed using a Leica TCS SPE confocal microscope under a 20× objective.

**DTR-mediated cell death and Chat conditional knockout assays.** After allowing 10–14 days for conditional viral expression (injections and viral construct described previously), mice were intraperitoneally (i.p.) injected 3 times daily for 5 days with 800 ng (4 ng µl<sup>-1</sup> working solution) of diphtheria toxin (Sigma, D0564) for optimal cell death of targeted cholinergic neurons. Female *Chat-cre<sup>+/-</sup>*; *R26<sup>LSL-tdTomato/+</sup>* mice were used initially to validate DTR-mediated cell death by visualizing DBB cholinergic cell loss. For remaining experiments, female *Chat-cre<sup>+/-</sup>* animals were used. For controls, age- and gender-matched *Chat-cre<sup>+/-</sup>* mice (stereotaxically injected identically into the DBB with AAV-FLEX-DTR-P2A-EYFP) were injected with equal volume (200 µl per injection) of sterile saline. Body weights and daily food intake were measured and averaged per group for each time point presented. For *Chat* conditional knockout assays, *Chat<sup>loxP/loxP</sup>* mice were injected bilaterally into the HDB as previously described with either AAV-EF1α-EGFP-P2A-CRE-WPRE-hGHPa for conditional knockout animals or AAV-EF1α-EGFP-WPRE-hGHPa for controls. Body weights and daily food intake were measured and averaged per group for each time point presented. For cell counts after *Chat* conditional knockout, 3 mice from each group were euthanized and brains were sectioned at 40 µm for ChAT immunohistochemistry. 10 sections representing the anterior, central, and posterior areas of the DBB were chosen, and blinded, total cell counts based on ChAT immunoreactivity were tallied. A count from all 10 sections from a single mouse brain were totalled and averaged for each group of 3 mice. Data were normalized to control levels of expression and represented as a mean percentage ± s.e.m.

**Activity monitoring and metabolic assays.** Activity (reported as time active per day), O<sub>2</sub> consumption, and metabolic blood assays were performed by the Baylor College of Medicine Mouse Metabolism Core before obesity phenotypes (3 days after diphtheria toxin (DT) treatment), during early stages of obesity and hyperphagia (3 weeks post-DT treatment), and at late stages of obesity and hyperphagia (3 months post-DT treatment). Activity and O<sub>2</sub> consumption assays were performed using the Oxymax Comprehensive Laboratory Animal Monitoring System (CLAMS, Columbus instruments). Lean mass and body fat content was assessed using quantitative MRI. Blood panel assays for cholesterol, leptin, insulin, and glucose were also performed by the Baylor College of Medicine Mouse metabolism core. Blood was collected via the tail vein. Mice were fasted for 4 h before measuring blood glucose.

**Paired feeding assays.** Paired feeding assays were performed with individually-housed, male DT-treated (DBB-ablated) and saline-treated (non-ablated) animals. Assays for determining the contribution of food intake on maintaining obesity were conducted on animals 12-weeks post-ablation. First, daily body weight and *ad libitum* food intake for all animals was recorded for 7 days to establish baselines for all animals. Then 1 control mouse and 1 experimental mouse were then randomly paired. All food from experimental cages was removed and only an equivalent amount of food consumed the previous day by a mouse's respective control partner was introduced to the cage. This restrictive period was done for 21 days. Afterwards, all experimental mice were allowed to resume to feed *ad libitum* once again and food intake and body weight were measured daily for 2 weeks. Change in body weight over time was normalized as a percentage of day 1 initial starting weight for each individual animal. For paired feeding conducted on animals used for *AgRP* and *Pomc* transcript analysis, assays were performed 3 days post-ablation to prevent significant weight gain from hyperphagia. A restrictive feeding period was conducted for 21 days, after which mice were euthanized on the morning after the final day, and hypothalamic tissue was harvested for RNA purification and subsequent qRT-PCR (see below).

**In vivo optogenetic behaviour assays.** Concurrent with AAV-EF1α-DIO-hChR2(H134R)-EYFP-WPRE-hGHPa injections (as described previously), male *Chat-cre<sup>+/-</sup>* mice were bilaterally implanted with 200 µm silica fibre optic implants made in-house (Thor Labs, TS1249968); 230 µm ferrules (Precision Fibre Products,



MM-FER2007C-2300) and situated 0.1 mm above the viral injection site. Fibre optic implants were held in place by a cap made from adhesive cement (C&B Metabond Quick! Cement System (Parkell)) for initial base, and crosslinked flash acrylic (Yates-Motloid, 44115 and 44119) for headcap. Mice were allowed at least two weeks for recovery and expression of the virus before assays were performed. For prolonged 2-day stimulation, each mouse was allowed 48 h to acclimate in a behaviour box with free access to food and water (days 1 and 2: acclimation). In addition, acclimation occurred while tethered to a dual fibre optic cord (Doric Lenses) attached to a 473 nm laser source (CrystaLaser CL-2005). After the 48-h acclimation period, a pre-measured amount of food was placed into the chamber and weighed once every 24 h for two days without stimulation (days 3 and 4: pre-stimulation). Over the next 48 h, food was weighed once each day while mice were chronically stimulated with trains of blue light (5 mW, 10 ms pulses, 20 Hz, 5 s trains, 30 s intervals) (days 5 and 6: stimulation). Finally, food was weighed once every 24 h for two final days with no blue light stimulation (days 7 and 8: post-stimulation). As a control group, non-ChR2-expressing mice were injected and implanted in the identical way used for experimental mice. Control mice were acclimated identically and were subsequently subjected to a mock stimulation for 48 h, and food intake was measured each day. For comparisons between pre-stimulation, stimulation, and post-stimulation conditions, paired Student's *t*-tests were used. For comparisons between experimental conditions and the control (mock-stimulation) condition, unpaired Student's *t*-tests were used. For short-term 2-h stimulation experiments, mice were given 48 h to acclimate in their behaviour chamber. After acclimation, mice were fasted overnight and subsequently presented with a pre-measured amount of food in the morning. For control conditions, mice were not stimulated and food intake was recorded every 30 min for 2 h total. For experimental conditions, mice were stimulated with trains of blue light (5 mW, 10 ms pulses, 20 Hz, 5 s trains, 30 s intervals) for 15 min before presentation of pre-measured chow, and food intake was recorded every 30 min for 2 h total in presence of continued blue light illumination. Trials were randomized and conducted one week apart on the same animals. For experiments targeted at terminal stimulation in the arcuate nucleus of the hypothalamus, male animals were bilaterally injected into the HDB with AAV-EF1 $\alpha$ -DIO-hChR2(H134R)-EYFP-WPRE-hGHpA, and a single fibre optic was implanted into the third ventricle at the level of the arcuate (from bregma: AP = -1.70, DV = -5.75, ML = 0.00). For behavioural assays, animals were first allowed 48 h to acclimate to their behaviour cage while tethered to a fibre optic cord. After acclimation, mice were fasted overnight. In the morning, mice were presented with standard pre-measured pellet chow and food intake was recorded every 30 min for 2 h total either under conditions of light stimulation (5 mW, 10 ms pulses, 20 Hz, 5 s trains, 30 s intervals) or no stimulation. Trials were randomized and conducted one week apart on the same animals. Paired statistics were used to compare 'stim' and 'no-stim' conditions on these animals. As a control group, Cre-negative male littermates were injected with virus and implanted in the same way as experimental mice. Behavioural assays on these mice were done the same way under conditions of light illumination. Comparisons between mock-stimulated and experimental cohorts were done using unpaired statistics. For terminal stimulation experiments in the presence of mecamylamine (Tocris, catalogue number 2843), mice were fasted overnight and presented with a pre-measured amount of chow in the morning. Mecamylamine was administered by i.p. injection at 1 mg per kg 15 min before the start of a 2-h feeding in the presence or absence of blue light illumination. Control (sterile 1  $\times$  PBS i.p. injections) and experimental trials were conducted 1 week apart.

**Electrophysiology and pharmacology of POMC-EGFP or NPY-hrGFP neurons, ChR2-expressing DBB neurons, and hM4D-expressing DBB neurons.** Brain slices containing the hypothalamic arcuate nucleus were prepared from 6–8-week-old *Pomc-EGFP*<sup>+/−</sup> or *Npy-hrGFP*<sup>+/−</sup> transgenic mice of either gender. For ChR2 stimulation and hM4D-mediated inhibition, brain slices containing the DBB were prepared from 12–16-week-old male animals expressing either ChR2::EYFP or hM4D-EGFP in the DBB, respectively. Animals were anaesthetized with isoflurane and brains were rapidly removed and transferred into sucrose-based cutting solution, containing (in mM): 250 sucrose, 25 NaHCO<sub>3</sub>, 1.25 NaH<sub>2</sub>PO<sub>4</sub>, 2.5 KCl, 1.5 MgCl<sub>2</sub>, 2 CaCl<sub>2</sub>, 10 glucose, and continuously bubbled with 5% CO<sub>2</sub> / 95% O<sub>2</sub>. 300- $\mu$ m-thick coronal brain slices were prepared using a Leica VT 1200 vibratome and placed for recovery in a 5% CO<sub>2</sub> / 95% O<sub>2</sub> bubbled regular ACSF solution, containing (in mM): 128 NaCl, 24 NaHCO<sub>3</sub>, 1 NaH<sub>2</sub>PO<sub>4</sub>, 3 KCl, 1 MgCl<sub>2</sub>, 1.6 CaCl<sub>2</sub>, 8 glucose. After at least 1-h recovery and 20–30 min before recording, slices were transferred into a recording chamber continuously perfused at 2 ml min<sup>−1</sup> with aforementioned ACSF at 24°C. POMC-, NPY-, ChR2::EYFP-, and hM4D-expressing neurons were identified by transmitted light DIC and EGFP fluorescent imaging using a Slicescope Pro 6000 optical setup (Scientifica), equipped with a CoolLED pE-100 470 nm excitation light source, 49002-ET-EGFP (FITC/Cy2) emission filter (Chroma Technology), and optiMOS camera (QImaging). Electrical activity of neurons was recorded in

whole-cell current clamp mode using a Multiclamp 700b amplifier and a 1440a Digidata interface (Molecular Devices). Pipette solution contained (in mM): 10 KCl, 120 K gluconate, 1 MgCl<sub>2</sub>, 10 HEPES, 1 EGTA, 5 Na<sub>2</sub>-ATP, 0.01 Na-GTP, pH 7.2. To test the cholinergic effects on POMC and NPY neurons, baseline neuronal activity was prerecorded for at least 10 min to assure a stable firing rate, after which acetylcholine (Sigma-Aldrich, A6625) was added to the bath perfusion at 100  $\mu$ M. For ChR2 stimulation, 12 consecutive trains of blue light were given at 30 s intervals. Each train lasted for 5 s with 10 ms light pulses delivered at 20 Hz. For recordings from hM4D-expressing cells, we recorded 90 s baseline sweeps and delivered 5 s current injections at 2 pA and 10 pA, spaced 25 s apart. This protocol was repeated again 6 min after CNO bath application. For all recordings, neuronal firing activity was analysed offline using the event detection feature of Clampfit 10.3 software (Molecular Devices). Repeated measures ANOVA with Holm–Sidak multiple comparison, and Sigma Plot 11.0 software (Systat Software) were used for statistical analyses of data, where applicable. For acute acetylcholine responses, a localized 2 s application of acetylcholine (100 mM, Sigma-Aldrich, A6625) was applied near a patched POMC-EGFP neuron (held at −70 mV) using a FemtoJet (Eppendorf). Each recording was performed using a 20-s sweep with an inter-trial interval of 1 min and repeated for 5 sweeps each for baseline, synaptic blockers (10  $\mu$ M CNQX (Tocris), 20  $\mu$ M APV (Tocris), 50  $\mu$ M GABazine (Tocris), and nicotinic blockers (10  $\mu$ M mecamylamine (Tocris), 0.1  $\mu$ M methyllycaconitine citrate (Tocris), and 10  $\mu$ M dihydro- $\beta$ -erythroidine hydrobromide (Tocris)).

**CAV-Cre tracing experiments.** *R26<sup>LSL-tdTomato</sup>* animals were stereotactically injected bilaterally with 70 nl of CAV-Cre virus (purchased from the vector core at the Institut de Génétique Moléculaire de Montpellier) targeted to the arcuate nucleus (from bregma: AP = -1.70, DV = -5.80, ML =  $\pm$ 0.20). Sections through the DBB were obtained and stained for ChAT using the identical ChAT IHC protocol detailed previously. All sections were mounted using DAPI Fluoromount-G (Southern Biotech, 0100-20). Detection of fluorescent expression was performed using a Leica TCS SPE confocal microscope under a 10 $\times$  or 20 $\times$  objective.

**hM4D-mediated feeding assays.** 12-week-old, male *Chat-cre*<sup>+/−</sup> mice were stereotactically injected bilaterally into the DBB (500 nl per hemisphere) with an hM4D-EGFP-expressing AAV (AAV-CBA-FLEX-hM4Di-P2A-EGFP-WPRE-sv40pA, serotype 2/9, Addgene, plasmid number 52536). After a two-week recovery, mice were fasted overnight in their home cage, and food was presented in the morning. Food intake was measured every 30 min for 2 h total, in the presence or absence of CNO. For control experiments, mice were injected i.p. with sterile saline and allowed to wait 15 min before food presentation. For experimental conditions, mice were injected i.p. with CNO (5 mg per kg) and allowed to wait 15 min before food presentation and measurement, which was recorded for 2 h in total. Trials were randomized and conducted one week apart on the same animals. Paired statistics were used to compare 'saline' and 'CNO' conditions on these animals.

**Transcript analyses.** For *Agrp* and *Pomc* transcript analysis from the arcuate nucleus, 4 DBB-ablated mice and 4 non-ablated mice were euthanized in the morning and their brains were immediately dissected. For AChR transcript analysis, 4 wild-type male mice were taken and brains dissected and processed identically. Small sections of the ventral hypothalamus containing the arcuate nucleus were dissected out and RNA was isolated following the TRIzol (Life Technologies, 15596-018) protocol for tissue homogenization and RNA isolation. In brief, tissue was placed in 1 ml of TRIzol reagent and homogenized using a 1.5 ml pre-sterilized pestle. Homogenized samples were allowed to incubate at room temperature for 5 min. 0.2 ml of chloroform was added, and the tube was shaken vigorously by hand for 15 s. The sample was again incubated at room temperature for 2 min and centrifuged at 12,000g for 15 min at 4°C. The upper aqueous phase was pipetted into a new tube, 0.5 ml of isopropanol was added, and the mixture was incubated at room temperature for 10 min. The tube was centrifuged at 12,000g for 10 min at 4°C. The supernatant was removed and the RNA pellet was washed with 1 ml of 75% ethanol. The sample was vortexed and centrifuged at 7,500g for 5 min at room temperature. The pellet was air-dried for 15 min and resuspended in 40  $\mu$ l RNase-free water at 60°C. RNA was DNase-digested using the manufacturer's protocol (Promega). DNase was inactivated via phenol-chloroform extraction. Purified RNA was quantified using a NanoDrop (Wilmington, DE), and first-strand cDNA was synthesized using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Negative controls did not contain reverse transcriptase. Transcripts were amplified using standard PCR conditions (95°C for 120 s, 95°C for 20 s, 60°C for 20 s, 72°C for 20 s, 34 cycles, 7°C for 300 s, 4°C until storage at −20°C). Amplified products were run on 2% agarose gels, imaged, and quantified using ImageJ software. AChR primer sequences were as follows (forward primer first, reverse primer second, in 5'–3' orientation), CHR1A1: GTCCAATAACGCCGCTGAGG, CTAGCGATGGCTATGGCTGG; CHR1A2: GACTCTTCGGTGAAGGAAGATTG, AGAGCAGAAGA TGGTTGTCCAG; CHR1A3: GCCAAAGAGATTCAAGATGATTGG, TCTGGGGCTATTGAGAAAGTGC; CHR1A4: GACTTCTCGGTGAAG

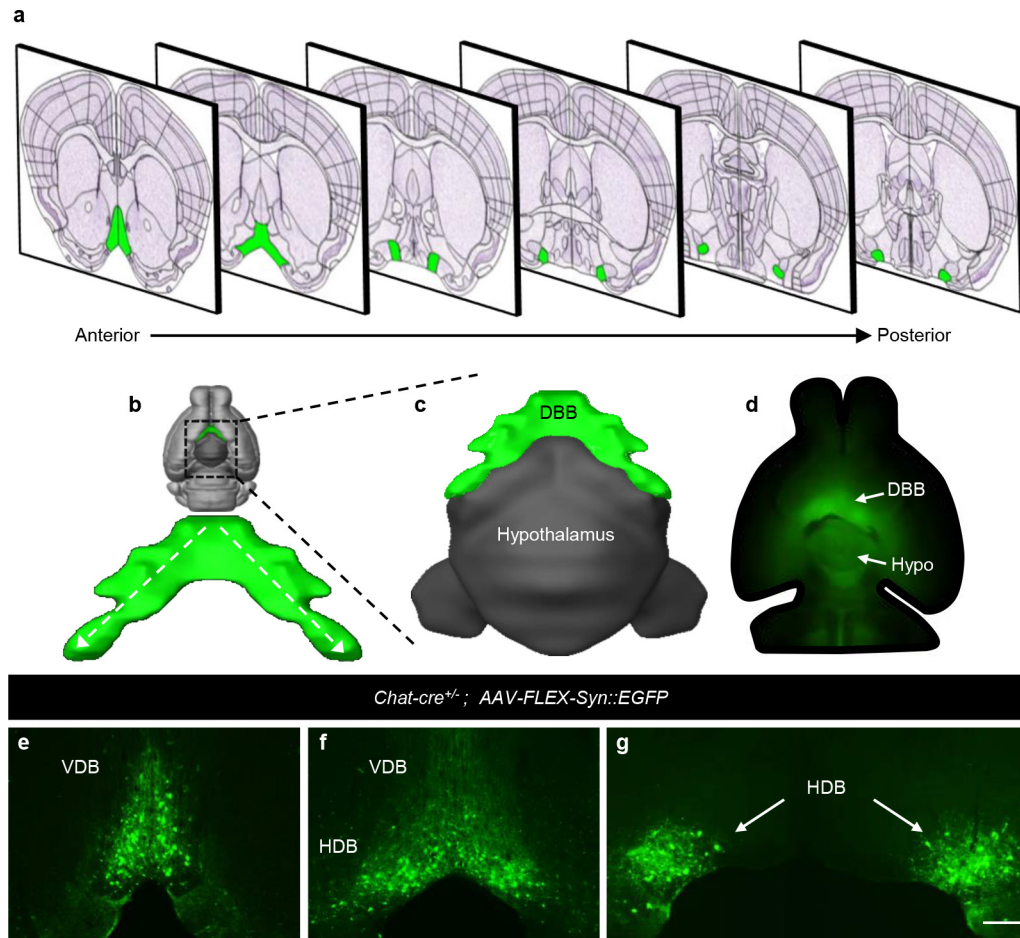
GAGGAC, GGAAGATGTGGGTGACTGACG; CHRNA5: CGTCCGCG AGGTTGTTGAAG, AGCTGCTTGACTGCTCACTAAG; CHRNA6: CAAACGAGGTATAAGACGACTG, TCTTGTGGGGCTAGCTCGG; CHRNA7: CCTAAGTGGACCAGGATCATTG, ATGTAGAGCAGGTTGCCATTGC; CHRNA9: GTCCCTCTGATAGGAAAATACTAC, CTAAGGCAGC TCTCACCCAC; CHRNA10: ACTCATCGGAAAGTACTATATGCG, GACTCTAATGGCTTGGACTGTC; CHRNB1: ACCAGATGCAGGAGAGA AGATG, GAGCGATGATGCAGGTTGAGG; CHRNB2: TGACCAGAGTG TGAGGGAGG, AGCTGCAAATGAGAGACCTCAC; CHRNB3: ACTTCATC AGTCAGGTTGTTCAAG, CTAGGTGGGATTCTCTCTATGTG; CHRNB4: ATCAGAGTGTATCGAGGACTG, CACTAGGCTGCTCATATCATCC; CHRM1: GCCAAGGTGATGCCCTTACTC, TGCCTGTCACTGTAGCCAGAG; CHRM2: AGAGCCCTGAAGTCGCAGATC, CTCCTGGATCTGGCTT TCAG; CHRM3: GGCTTCCTGGCATTGGTGAC, GCCAGAGGTCACAGG CTAAG; CHRM4: TGACTGGTTCCCTGAGCCTG, AGTAGCCCTTGATGAT GTATAAGG; CHRM5: ACTATTACCTGCTCAGCTTGCG, GTAACGATCAA AGCTAATCACCAAG.

AgRP products were quantified by qPCR using the following primer pair: GCGGAGGTGCTAGATCCACAGAA and AGGACTCGTCAGCCTTACAC. POMC products were quantified by qPCR using the following primer pair: AGAACGCCATCATCAAGAAC and AAGAGGCTAGAGGTCATCAG. Actin was used as a reference control using the following primer pair: GCAAGCAGGAGTACGATGAG and TAACAGTCCGCCTAGAAGCA. For quantitative transcript analysis, all reactions were done in triplicates with no reverse

transcriptase negative control samples. Samples were prepared according to the BIO-RAD iQ SYBR Green Supermix instructions. Briefly, cDNA was diluted to  $1 \text{ ng} \mu\text{l}^{-1}$  and primers were diluted to  $2.5 \mu\text{M}$  stock of combined forward and reverse primers. Reactions were set up in  $15 \mu\text{l}$  total volume per sample ( $7.5 \mu\text{l}$  SYBR Green Supermix,  $5 \mu\text{l}$  cDNA,  $1.5 \mu\text{l}$  forward and reverse primer mix,  $1 \mu\text{l}$  water) in Applied Biosystems Advanced Studio 3 compatible 96-well plates. Plates were sealed with adhesive film and mixed thoroughly before amplification. Amplification occurred on the Applied Biosystems Advanced Studio 3 qPCR machine on its standard amplification protocol. All transcript analysis was standardized to the amplification curve of actin for each sample, and a Student's *t*-test was performed to analyse differences in transcript expression among samples.

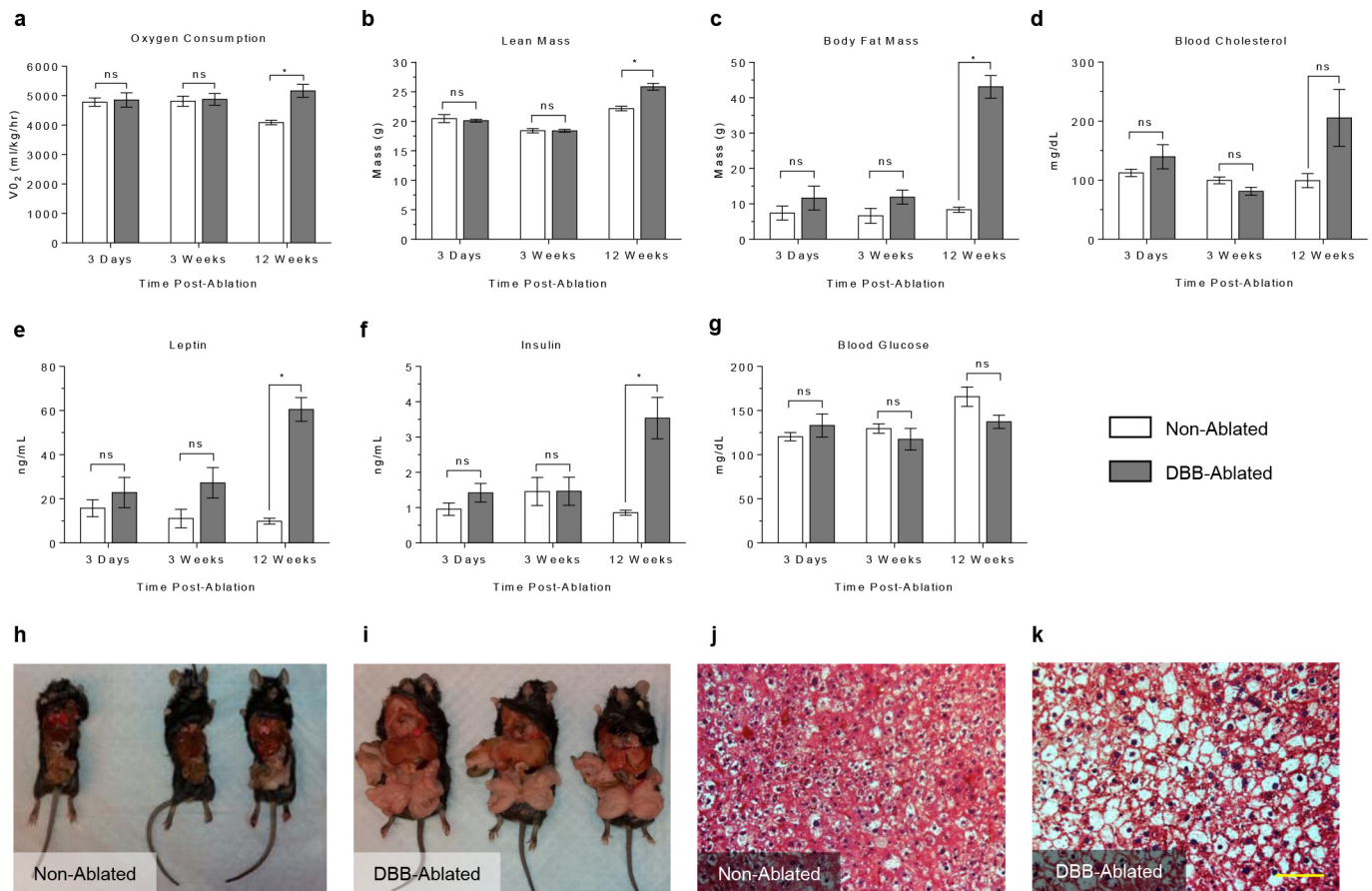
**Statistical analyses.** With the exception of electrophysiological-based experiments (previously described), all other statistical analyses were performed using GraphPad Prism 6 software (GraphPad), accounting for appropriate distribution and variance to ensure proper statistical parameters were applied. Experimental sample sizes were chosen according to minimal accepted norms within the field. With regards to experimental randomization for cell ablation, *Chat* conditional knockout, and optogenetic behaviour assays, mice were randomly separated into two groups before manipulation by an independent, blinded assistant not involved in experimentation or experimental design. For quantification of *Chat* conditional knockout in the DBB, DBB images were acquired by an independent assistant not involved in the experimentation and cell counts were then objectively tallied by a second assistant without knowledge of the experimental groups. Statistical methods used are described in figure legends for the respective experiments.





**Extended Data Figure 1 | Anatomy of the diagonal band of Broca.** **a–c**, The DBB is situated in the basal forebrain and consists of an anteriorly located vertical limb (VDB), which branches posteriorly into separate bilateral horizontal limbs (HDB). **d**, *AAV-FLEX-tdTomato* (pseudocoloured in green) injected bilaterally into the HDB to show

expression throughout the full extent of the diagonal band (VDB and HDBs). **e–g**, Bilateral viral injections into the HDB (*AAV-FLEX-Syn::EGFP* shown as an example) are sufficient to target the full extent of the diagonal band (VDB and HDB). Scale bars, 100  $\mu\text{m}$ . Schematic in **a** was adapted from images from the Allen Brain Institute Reference Atlas<sup>30</sup>.

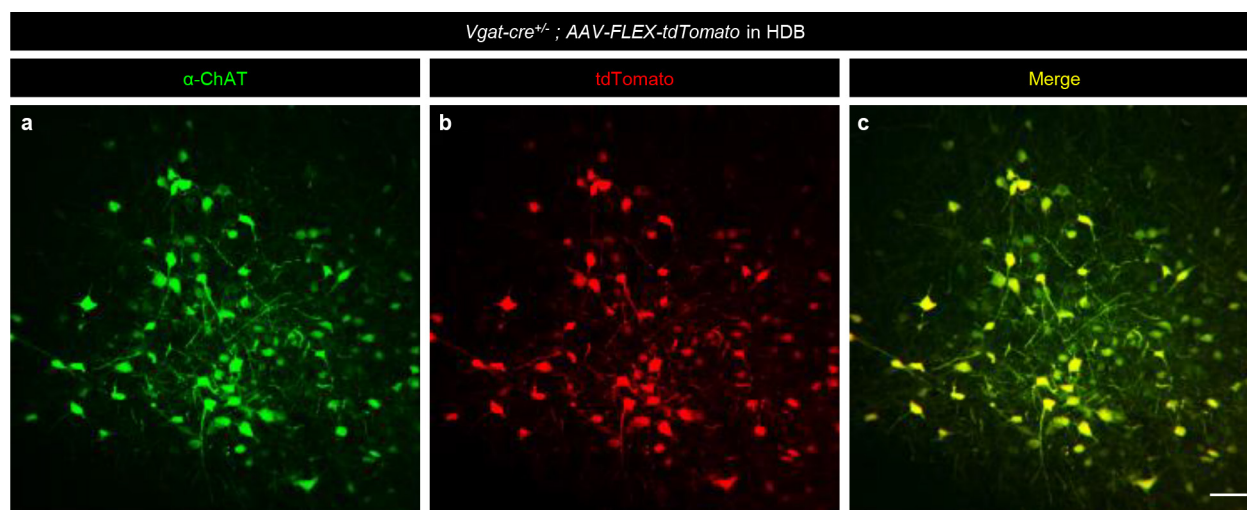


**Extended Data Figure 2 | Time-dependent changes in body content and metabolic measures between non-ablated and DBB-ablated mice.**

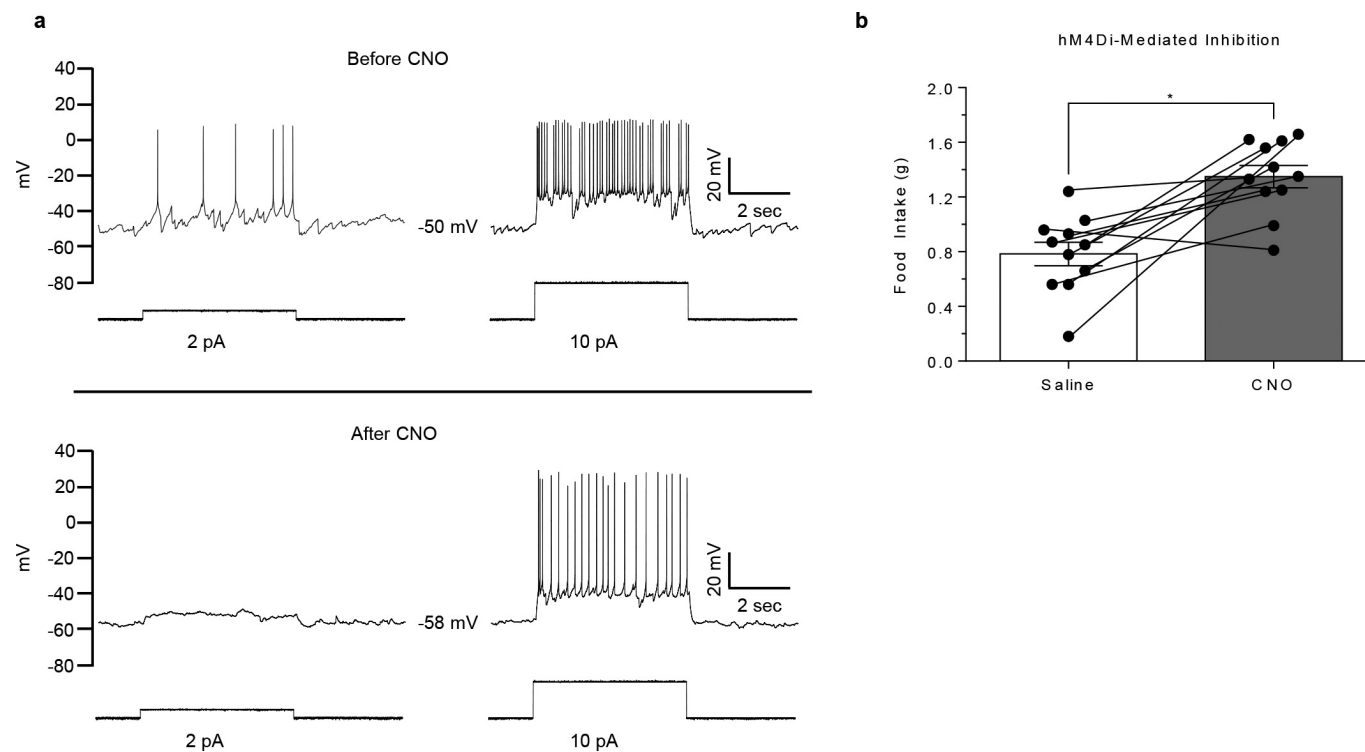
**a**, Oxygen consumption, represented as VO<sub>2</sub>, at various stages post DBB-ablation. Data are represented as mean ± s.e.m. \**P* < 0.05 by two-sided, unpaired Student's *t*-test. **b**, **c**, Average total lean mass (**b**) or body fat (**c**) content at various stages post DBB-ablation. Data are represented as mean ± s.e.m. \**P* < 0.05 by two-sided, unpaired Student's *t*-test. **d**–**f**, Blood cholesterol (**d**), leptin (**e**), insulin (**f**), and blood glucose concentration (**g**)

at various stages post DBB-ablation. Data are represented as mean ± s.e.m. \**P* < 0.05 by two-sided, unpaired Student's *t*-test.

**h**, **i**, Representative non-ablated (**h**) or DBB-ablated (**i**) animals (*n* = 3 mice) dissected to show abdominal fat pads and enlarged fatty livers. **j**, **k**, Representative haematoxylin and eosin (H&E) stains showing fat deposition of the liver in non-ablated (**j**) or DBB-ablated (**k**) animals. Scale bar, 100 μm.



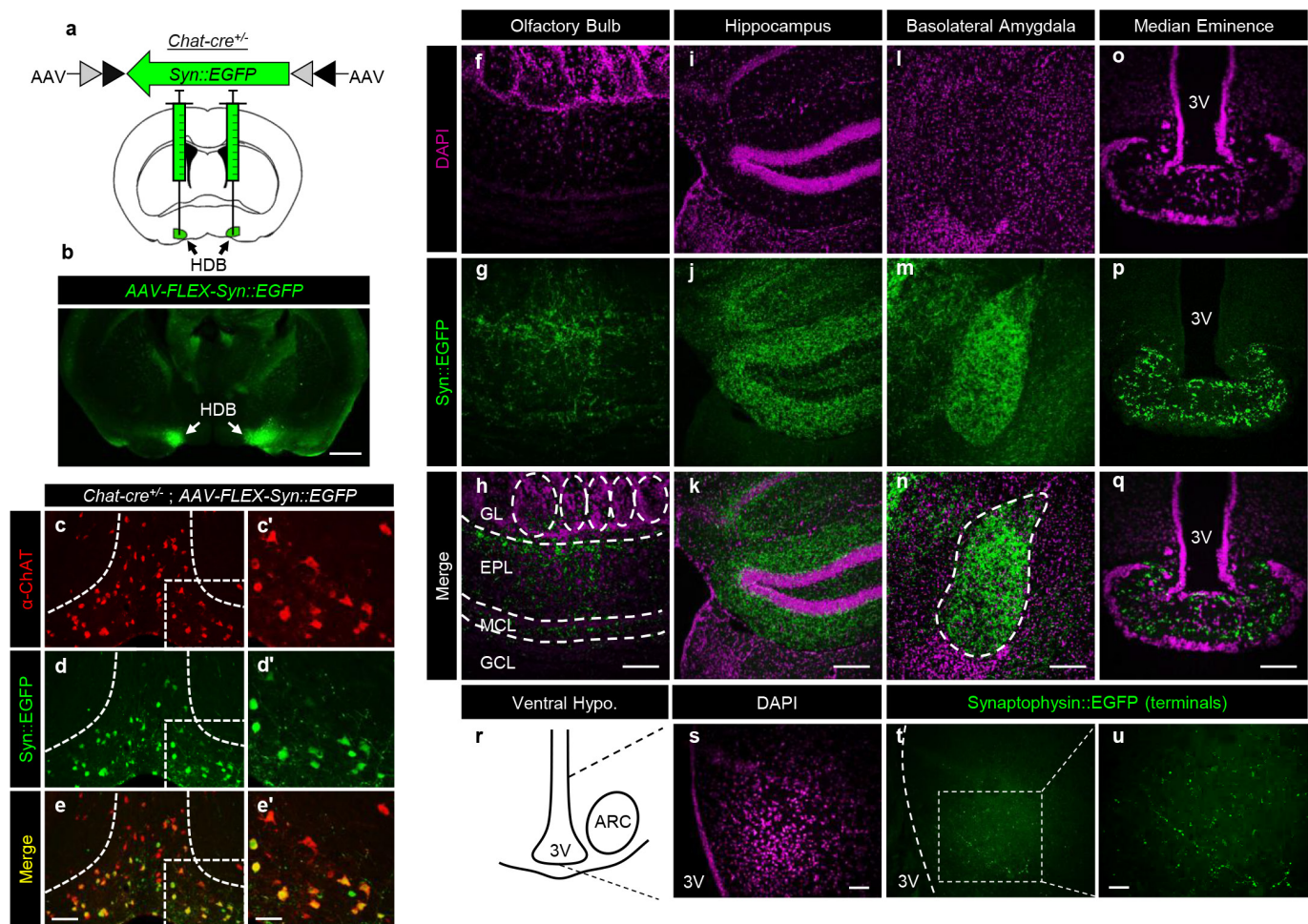
**Extended Data Figure 3 | Cholinergic DBB neurons are VGAT-positive.** a–c, *Vgat-cre<sup>+/+</sup>* mouse injected with AAV-FLEX-tdTomato into the HDB and stained with anti-ChAT antibody (green channel). Data suggest a high percentage of co-localization between VGAT-positive neurons and ChAT. Scale bar, 100  $\mu$ m.



**Extended Data Figure 4 | hM4D-mediated inhibition of the cholinergic DBB increases food intake.** **a**, Whole-cell electrophysiological recordings from a representative cholinergic neuron expressing the hM4D inhibitory DREADD receptor before (top) and after (bottom) CNO treatment.

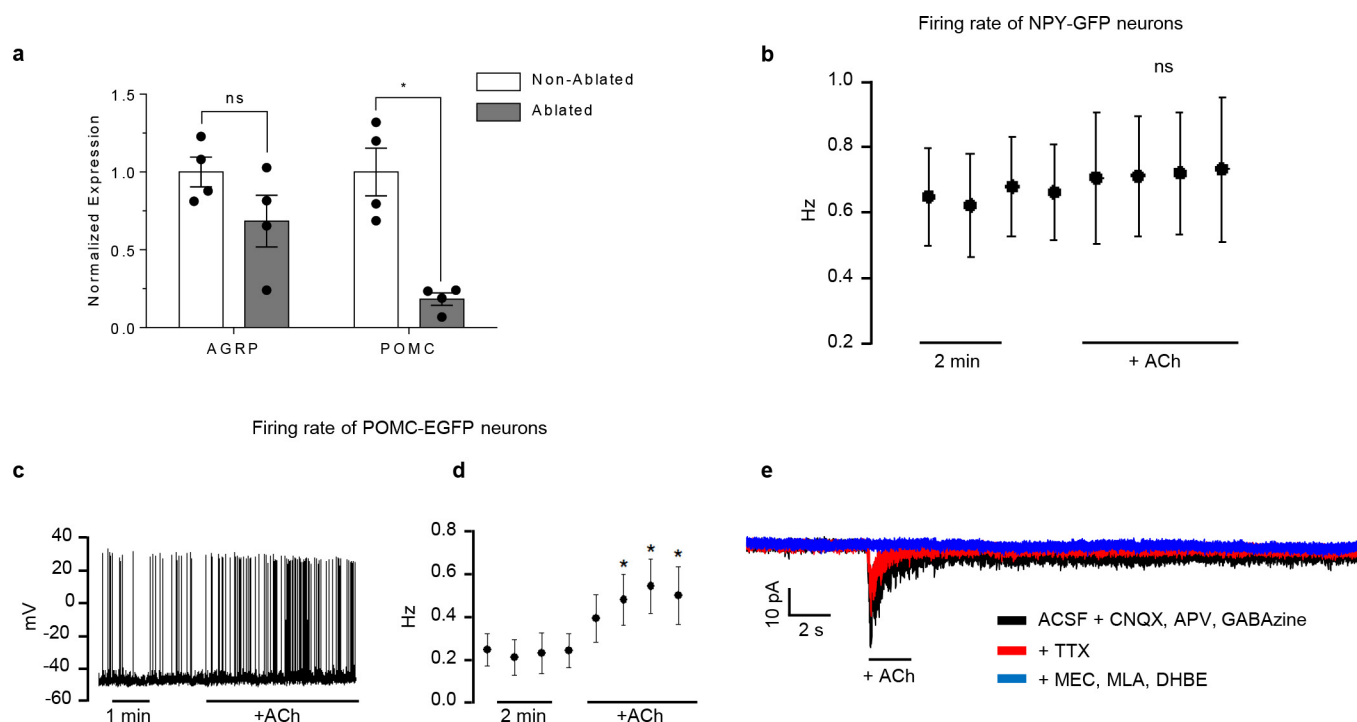
**b**, Total two-hour food intake after an overnight fast ( $n = 11$  mice). 'Saline' and 'CNO' groups represent separate trials of the same animals in the absence or presence of CNO, respectively. Data are represented as mean  $\pm$  s.e.m.  $*P < 0.05$  by two-sided, paired Student's  $t$ -test.





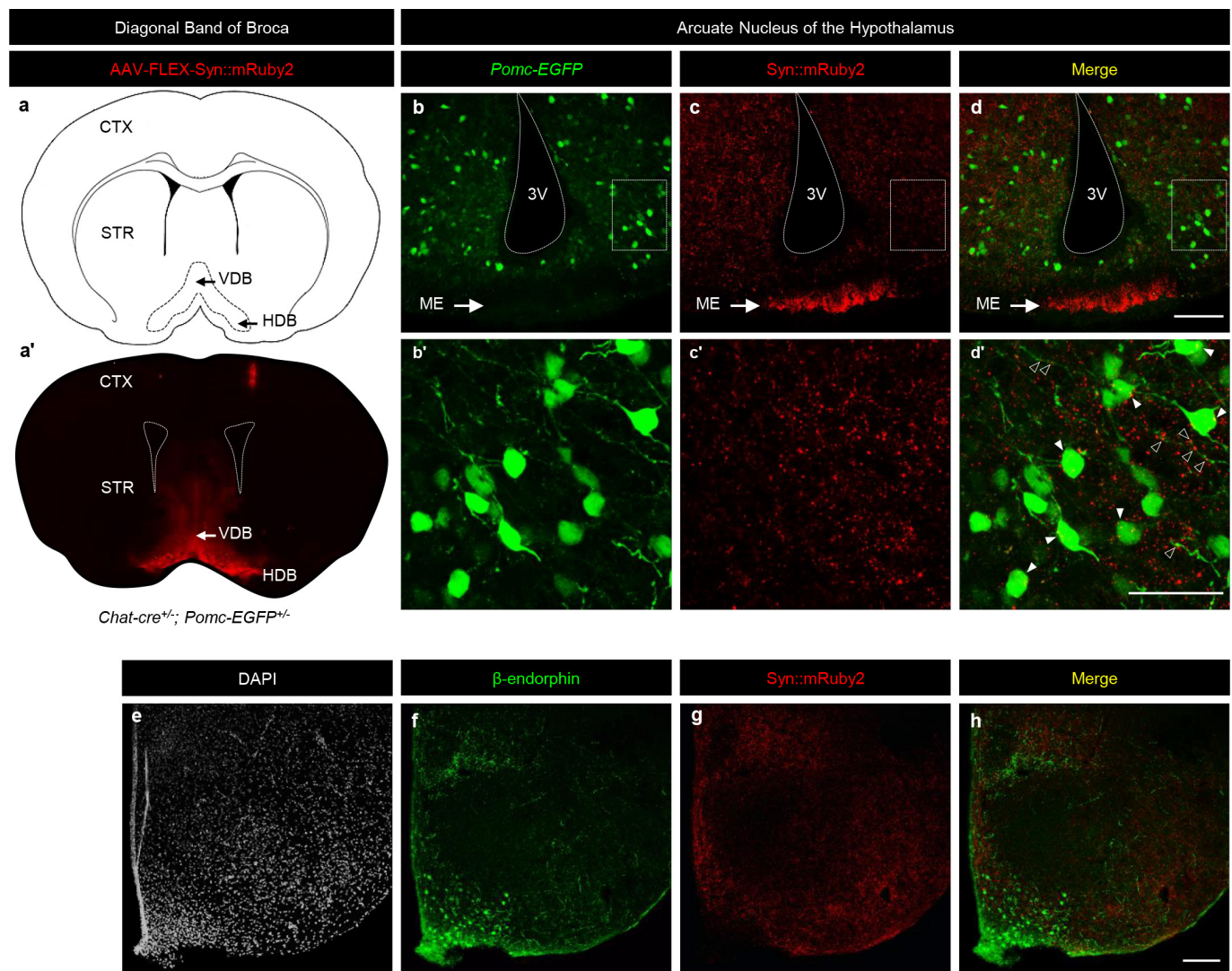
**Extended Data Figure 5 | Representative cholinergic DBB-innervating projection sites in the brain.** **a, b**, Schematic (**a**) and representative image (**b**) showing injection site in the HDBs of a *Chat-cre<sup>+/-</sup>* mouse with AAV-FLEX-Syn::EGFP. Scale bar, 1 mm. **c-e**, *Chat-cre<sup>+/-</sup>* mouse injected bilaterally into the HDB with AAV-FLEX-Syn::EGFP and stained with an anti-ChAT antibody. Staining shows high co-localization of ChAT-positive neurons with the Cre-dependent AAV, demonstrating the efficacy of the *Chat-cre* mouse line. Of the neurons infected with virus, only a small fraction do not express ChAT. Scale bars, 100  $\mu$ m (**c-e**) and 50  $\mu$ m

(**c'-e'**). **f-h**, Cholinergic projections from the DBB can be validated by known projection sites including the olfactory bulb. Scale bar, 200  $\mu$ m; GL, glomerular layer; EPL, external plexiform layer; MCL, mitral cell layer; GCL, granule cell layer. **i-k**, Hippocampus. Scale bar, 200  $\mu$ m. **l-n**, Amygdala (basolateral amygdala shown here). Scale bar, 100  $\mu$ m. **o-u**, Also shown are areas of the ventral hypothalamus including, Median eminence (**o-q**), and arcuate nucleus (**r-u**). Scale bars, 100  $\mu$ m. Schematic in **a** was adapted from an image from the Allen Brain Institute Reference Atlas<sup>30</sup>.



**Extended Data Figure 6 | *Pomc* expression is reduced in pair-fed DBB-ablated animals.** **a**, Relative *AgRP* and *Pomc* expression levels in arcuate nuclei between pair-fed non-ablated and DBB-ablated mice ( $n = 4$  mice per group). Data are represented as mean  $\pm$  s.e.m.  $*P < 0.05$  by two-sided, unpaired Student's *t*-test. **b**, Acetylcholine does not significantly alter arcuate NPY neuron firing. Data shown as mean  $\pm$  s.e.m. of spike frequency in 8 acetylcholine-treated cells (1 recorded neuron per slice, 1–2 slices per mouse,  $n = 6$  mice), comparisons made versus baseline values (1–4 min) by repeated measures ANOVA with Holm–Sidak multiple comparison. **c**, **d**, Acetylcholine significantly increases arcuate POMC neuron firing. Data shown as mean  $\pm$  s.e.m. of spike frequency in 7 acetylcholine-treated cells (1 recorded neuron per slice, 1–2 slices

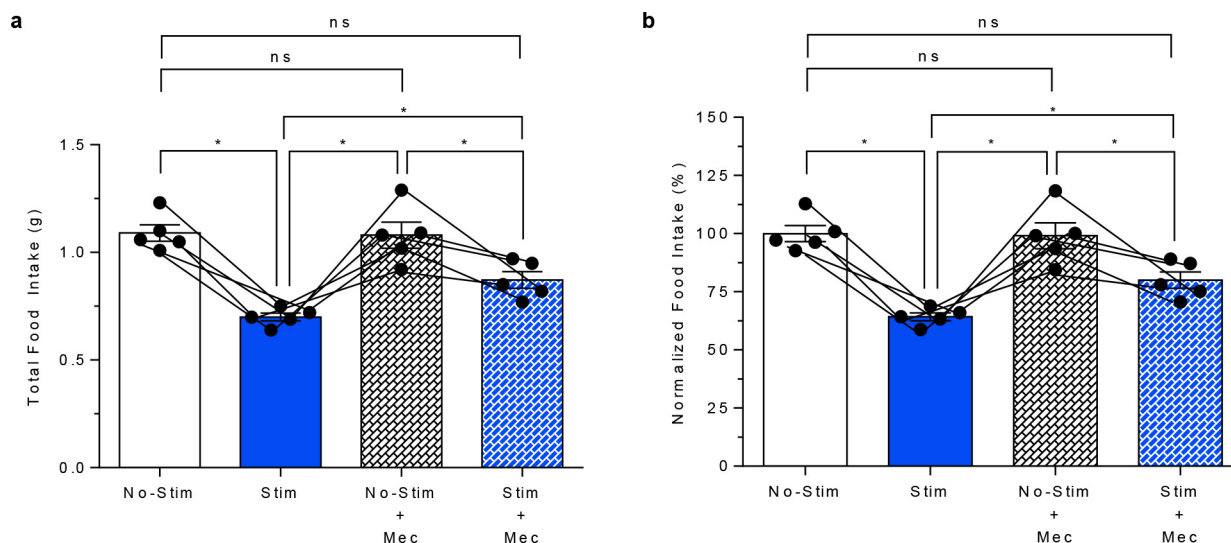
per mouse,  $n = 4$  mice),  $*P < 0.001$  versus baseline values (1–4 min) by repeated measures ANOVA with Holm–Sidak multiple comparison. **e**, Representative voltage clamp recording from POMC-EGFP neuron after local administration of acetylcholine and after pharmacological manipulation of fast synaptic transmission, as well as nicotinic acetylcholine receptor blockade (20-s sweep with an inter-trial interval of 1 min and repeated for 5 sweeps each for baseline, synaptic blockers (CNQX, 2-amino-5-phosphonopentanoic acid (APV), GABA and tetrodotoxin (TTX)), and nicotinic blockers (mecamylamine (MEC), methyllycaconitine (MLA) and dihydro- $\beta$ -erythroidine hydrobromide (DBHE)).



**Extended Data Figure 7 | Diffuse cholinergic DBB projections into the hypothalamus.** **a**, Site of AAV-FLEX-Syn::mRuby2 injections in the DBB. **b–d**, Representative images of cholinergic terminals (red) from the DBB into the hypothalamus. Scale bar, 100  $\mu$ m. ME, median eminence. **b'–d'**, Higher magnification images from the section shown in the dotted squares in **b–d**. Scale bar, 50  $\mu$ m. Closed arrowheads indicate close apposition or co-localization between cholinergic terminals and POMC (green) cell bodies, whereas open arrowheads show apposition or

co-localization on or near POMC neuronal processes. Of note, cholinergic terminals also appear on non-POMC neurons. **e, f**, DAPI (**e**), anti- $\beta$ -endorphin (**f**) (POMC neuron marker), **g, h**, AAV-FLEX-Syn::mRuby2 (injected into the DBB) (**g**), and merged (**h**) channels from **f** and **g** show innervation of the hypothalamus from cholinergic neurons in the DBB. Scale bar, 200  $\mu$ m. Schematic in **a** was adapted from an image from the Allen Brain Institute Reference Atlas<sup>30</sup>.

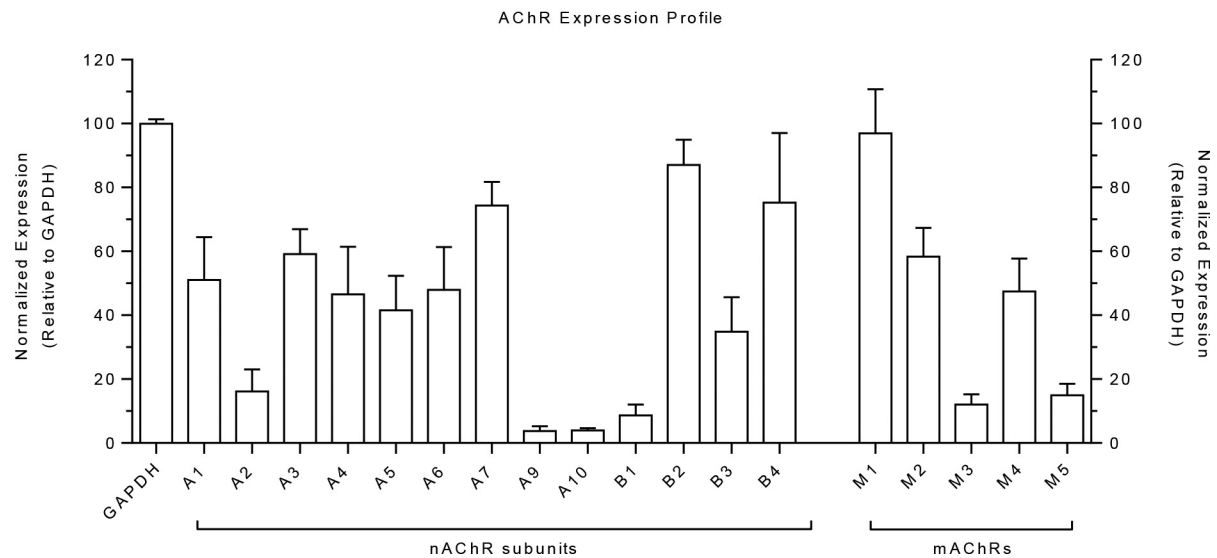




**Extended Data Figure 8 | ChR2-mediated decrease in feeding is partially suppressed by nAChR blockade. a, b,** Total (a) and normalized (with respect to the mean 'No-stim' value; b) food intake after a two-hour re-feeding after an overnight fast under specified conditions. Comparisons

between 'No-Stim,' 'Stim,' 'No-Stim + Mec,' and 'Stim + Mec' refer to separate trials conducted on the same animals ( $n = 5$  mice). Data are represented as mean  $\pm$  s.e.m. \* $P < 0.05$  by two-sided, paired Student's  $t$ -test.



**a**

**Extended Data Figure 9 | Arcuate acetylcholine receptor (AChR) expression profile. a,** Region-specific transcript expression profile of nicotinic and muscarinic AChRs in the arcuate nucleus, relative to the housekeeping gene, *Gapdh* ( $n = 4$  mice). Data are represented as mean  $\pm$  s.e.m.

# Evidence for a limit to human lifespan

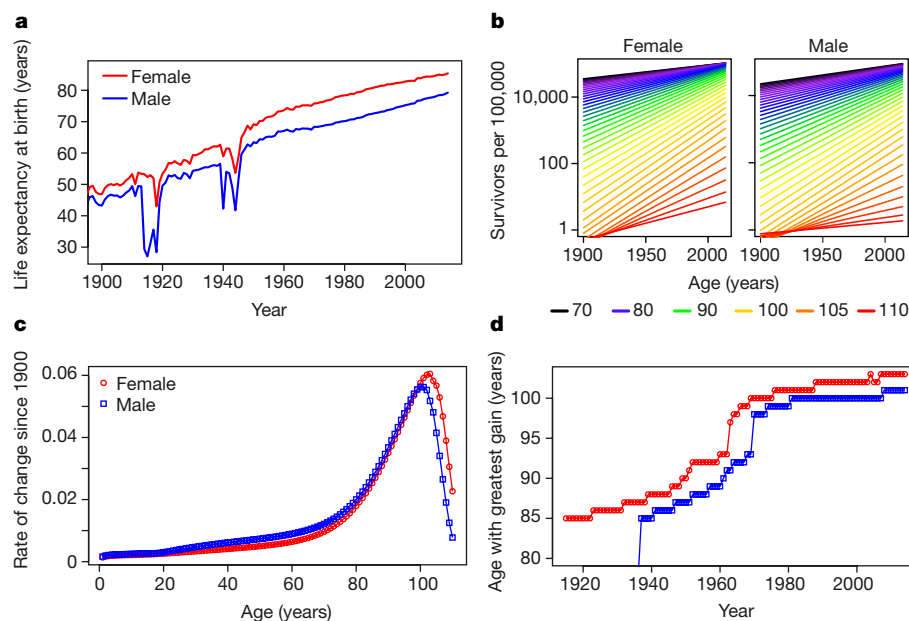
Xiao Dong<sup>1\*</sup>, Brandon Milholland<sup>1\*</sup> & Jan Vijg<sup>1,2</sup>

Driven by technological progress, human life expectancy has increased greatly since the nineteenth century. Demographic evidence has revealed an ongoing reduction in old-age mortality and a rise of the maximum age at death, which may gradually extend human longevity<sup>1,2</sup>. Together with observations that lifespan in various animal species is flexible and can be increased by genetic or pharmaceutical intervention, these results have led to suggestions that longevity may not be subject to strict, species-specific genetic constraints. Here, by analysing global demographic data, we show that improvements in survival with age tend to decline after age 100, and that the age at death of the world's oldest person has not increased since the 1990s. Our results strongly suggest that the maximum lifespan of humans is fixed and subject to natural constraints.

Maximum lifespan is, in contrast to average lifespan, generally assumed to be a stable characteristic of a species<sup>3</sup>. For humans, the maximum reported age at death is generally set at 122 years, the age at death of Jeanne Calment, still the oldest documented human individual who ever lived<sup>4</sup>. However, some evidence suggests that maximum lifespan is not fixed. Studies in model organisms have shown that maximum lifespan is flexible and can be affected by genetic and pharmacological interventions<sup>5</sup>. In Sweden, based on a long series of reliable information on the upper limits of human lifespan, the

maximum reported age at death was found to have risen from about 101 years during the 1860s to about 108 years during the 1990s<sup>6</sup>. According to the authors, this finding refutes the common assertion that human lifespan is fixed and unchanging over time<sup>6</sup>. Indeed, the most convincing argument that the maximum lifespan of humans is not fixed is the ongoing increase in life expectancy in most countries over the course of the last century<sup>1,2</sup>. Figure 1a shows this increase for France, a country with high-quality mortality data, but very similar patterns were found for most other developed nations (Extended Data Fig. 1). Hence, the possibility has been considered that mortality may decline further, breaking any pre-conceived boundaries of human lifespan<sup>1,7</sup>.

As shown by data from the Human Mortality Database<sup>8</sup>, many of the historical gains in life expectancy have been attributed to a reduction in early-life mortality. More recent data, however, show evidence for a decline in late-life mortality, with the fraction of each birth cohort reaching old age increasing with calendar year. In France, the number of individuals per 100,000 surviving to old age (70 and up) has increased since 1900 (Fig. 1b), which points towards a continuing increase in human life expectancy. This pattern is very similar across the other 40 countries and territories included in the database (Extended Data Figs 2, 3). However, the rate of improvement in survival peaks and then declines for very old age levels (Fig. 1c), which points

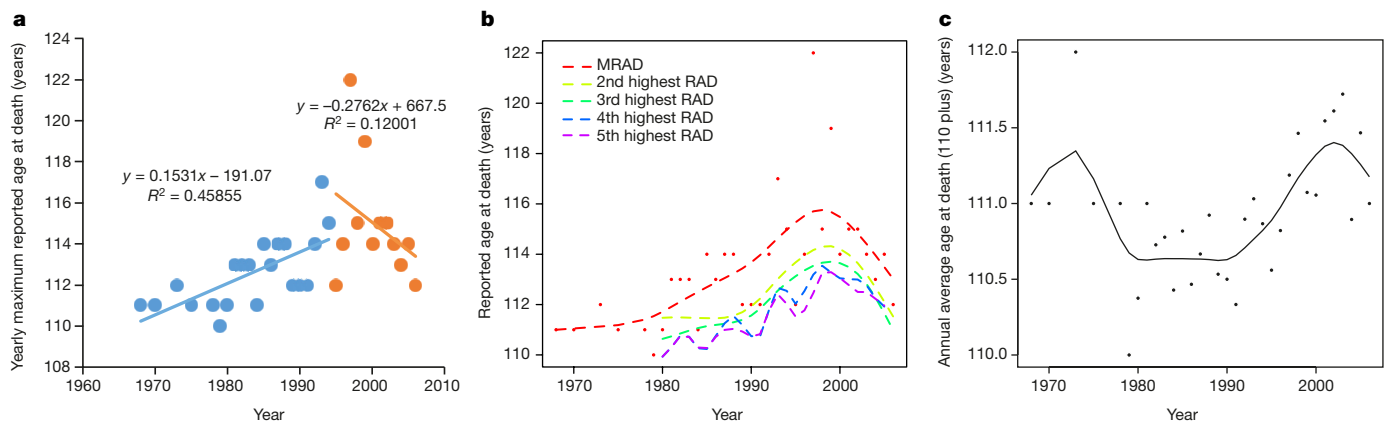


**Figure 1 | Trends in life expectancy and late-life survival.** **a**, Life expectancy at birth for the population in each given year. Life expectancy in France has increased over the course of the 20th and early 21st centuries. **b**, Regressions of the fraction of people surviving to old age demonstrate that survival has increased since 1900, but the rate of increase appears to be slower for ages over 100. **c**, Plotting the rate of

change (coefficients resulting from regression of log-transformed data) reveals that gains in survival peak around 100 years of age and then rapidly decline. **d**, Relationship between calendar year and the age that experiences the most rapid gains in survival over the past 100 years. The age with most rapid gains has increased over the century, but its rise has been slowing and it appears to have reached a plateau.

<sup>1</sup>Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA. <sup>2</sup>Department of Ophthalmology & Visual Sciences, Albert Einstein College of Medicine, Bronx, New York 10461, USA.

\*These authors contributed equally to this work.



**Figure 2 | Reported age at death of supercentenarians.** All data were collected from the IDL database (France, Japan, UK and US, 1968–2006). **a**, The yearly maximum reported age at death (MRAD). The lines represent the functions of linear regressions. **b**, The annual 1st to 5th highest reported ages at death (RAD). The dashed lines are estimates of the RAD using cubic

smoothing splines. The red dots represent the MRAD. **c**, Annual average age at death of supercentenarians (110 years plus,  $n = 534$ ). The solid line is the estimate of the annual average age at death of supercentenarians, using a cubic smoothing spline.

towards diminishing gains in reduction of late-life mortality and a possible limit to human lifespan. The same pattern was found across other developed, low-mortality countries (Extended Data Fig. 4). However, we considered the possibility that the age experiencing the greatest increase in survivorship increases with calendar years; that is, the peak in the rate of increase in survivorship to old age will shift to the right over time. To test this, we plotted the age at which this peak occurred against calendar years (Fig. 1d). The results indicate that the age with greatest improvement in survival appeared to plateau around 1980. A similar pattern was seen in 88% of the 41 countries in the database (Extended Data Fig. 5). Together, these findings suggest, but do not prove, that human lifespan may have a natural limit. To further investigate this idea, we turned our attention from late-life mortality to maximum human lifespan itself and examined the ages at death of the world's oldest individuals.

We first plotted the yearly maximum reported age at death (MRAD) of France, Japan, UK and US, countries with the largest number of recorded supercentenarians (individuals aged 110 years old or more) in the International Database on Longevity<sup>9</sup> (IDL;  $n = 534$ , 1968–2006). As shown in Fig. 2a, although age at death increased rapidly between the 1970s and early 1990s, it reached a plateau around 1995, close to 1997, the year Jeanne Calment died. We partitioned the data into two groups (1968–1994 and 1995–2006) and modelled each group using linear regression. The results indicate a trend break between the two groups. Before 1995, the MRAD increased by 0.15 years per year ( $r = 0.68$ ,  $P = 0.0007$ ); however, after 1995 it no longer increased significantly and in fact decreased slightly by 0.28 years per year ( $r = -0.35$ ,  $P = 0.27$ ). When we considered MRAD records from another, independent resource, the Gerontological Research Group (GRG; <http://www.grg.org/>), we observed a similar trend—an increase by 0.12 years per year ( $r = 0.71$ ,  $P = 0.0002$ ) during the period 1972–1994, followed by a slight decrease by 0.14 years per year ( $r = -0.36$ ,  $P = 0.70$ ) during the period 1995–2015 (Extended Data Fig. 6). These results indicate that although the MRAD increased until the 1990s, no further increases were observed after that time; human yearly MRAD has plateaued at 114.9 (95% CI: 113.1–116.7) years. To approximate the absolute limit of human lifespan, we modelled the MRAD as a Poisson distribution; we found that the probability of an MRAD exceeding 125 in any given year is less than 1 in 10,000.

One potential confounder of our results is the fairly small number of reported MRAD cases, which could explain these results simply as fluctuations. To provide a robust statistical model that would strengthen the observed pattern, we considered several series of high reported age at death (HRAD), that is, the highest RAD (MRAD) and the second to the fifth highest RADs (Fig. 2b; data summarized from IDL).

All series showed the same pattern as the MRAD. Notably, even the annual average age at death for these supercentenarians has not increased since 1968 (Fig. 2c).

Hence, in contrast to previous suggestions that human longevity can be extended ever further<sup>1</sup>, our data strongly suggest that the duration of life is limited. In the past, others have suggested that human lifespan is limited. For example, in 1980 Fries argued that increased prevention of premature deaths would lead to a compression of morbidity owing to a finite lifespan<sup>10</sup>. However, his arguments for such a limit to life, that is, the lack of a detectable increase in centenarians or in the maximum reported age at death, while correct at that time, have been refuted since<sup>2,6</sup>. Ten years later, Olshansky *et al.*<sup>11</sup> estimated the upper limits to human longevity based on hypothetical reductions in mortality rates, concluding that life expectancy at birth would not exceed 85 years. Like Fries, Olshansky *et al.* also suggested a biological limit to life based on the lack of an increase in the age of the verified longest-lived individual. However, as they mention, insufficient data prevented them from drawing definite conclusions. Now, more than two decades later, such data are becoming available. With the caveat that the ages at death of the supercentenarians in our present study are still noisy and made up of small samples, we feel that the observed trajectories in Fig. 2 are compelling and our results strongly suggest that human lifespan has a natural limit.

What could be the biological causes of this limit to human lifespan? The idea that ageing is a purposeful, programmed series of events that evolved under the direct force of natural selection to cause death has now been all but discredited<sup>12</sup>. Instead, what appears to be a 'natural limit' is an inadvertent byproduct of fixed genetic programs for early life events, such as development, growth and reproduction. Limits to the duration of life could well be determined by a set of species-specific, longevity-assurance systems encoded in the genome that counteract these inadvertent byproducts, which are likely to include inherent imperfections in transferring genetic information into cellular function<sup>13,14</sup>. To further extend human lifespan beyond the limits set by these longevity-assurance systems would require interventions beyond improving health span, some of which are currently under investigation<sup>15</sup>. Although there is no scientific reason why such efforts could not be successful, the possibility is essentially constrained by the myriad of genetic variants that collectively determine species-specific lifespan<sup>16</sup>.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 29 May; accepted 25 August 2016.**

**Published online 5 October 2016.**

1. Oeppen, J. & Vaupel, J. W. Demography. Broken limits to life expectancy. *Science* **296**, 1029–1031 (2002).
2. Vaupel, J. W. Biodemography of human ageing. *Nature* **464**, 536–542 (2010).
3. Austad, S. N. in *Molecular and Cellular Biology of Aging* (eds Vijg, J., Campisi, J. & Lithgow, G.) Ch. 2 (The Gerontological Society of America, 2015).
4. Jeune, B. *et al.* in *Supercentenarians* (eds H. Maier *et al.*) (Springer, 2010).
5. Kenyon, C. The plasticity of aging: insights from long-lived mutants. *Cell* **120**, 449–460 (2005).
6. Wilmoth, J. R., Deegan, L. J., Lundström, H. & Horiuchi, S. Increase of maximum life-span in Sweden, 1861–1999. *Science* **289**, 2366–2368 (2000).
7. Blagosklonny, M. V. Why human lifespan is rapidly increasing: solving “longevity riddle” with “revealed-slow-aging” hypothesis. *Aging* **2**, 177–182 (2010).
8. *The Human Mortality Database* (<http://www.mortality.org>, 2016).
9. Maier, H. *et al.* *Supercentenarians* (Springer, 2010).
10. Fries, J. F. Aging, natural death, and the compression of morbidity. *N. Engl. J. Med.* **303**, 130–135 (1980).
11. Olshansky, S. J., Carnes, B. A. & Cassel, C. In search of Methuselah: estimating the upper limits to human longevity. *Science* **250**, 634–640 (1990).
12. Vijg, J. & Kennedy, B. K. The Essence of Aging. *Gerontology* **62**, 381–385 (2016).
13. Finch, C. E. *Longevity, Senescence, and the Genome* (Univ. Chicago Press, 1990).
14. Vijg, J. *Aging of the Genome* (Oxford, 2007).
15. Longo, V. D. *et al.* Interventions to slow aging in humans: are we ready? *Aging Cell* **14**, 497–510 (2015).
16. Vijg, J. & Campisi, J. Puzzles, promises and a cure for ageing. *Nature* **454**, 1065–1071 (2008).

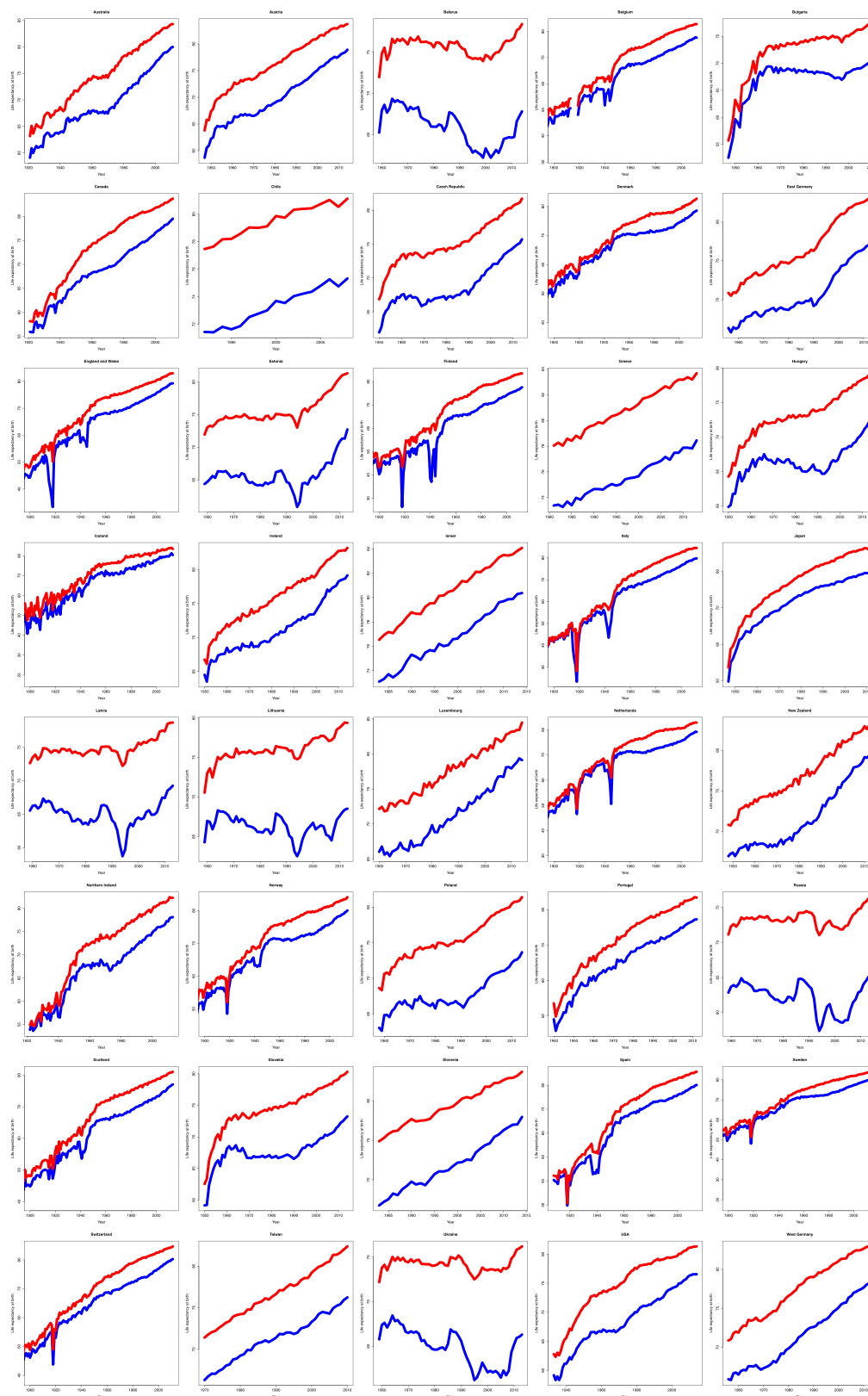
**Acknowledgements** We thank T. Wang for his suggestions on the statistical analysis. This study was supported by grants from the NIH to J.V. (AG017242 and AG047200), the Albert Einstein College of Medicine Institute for Aging Research/Nathan Shock Center, and the Paul F. Glenn Center for the Biology of Human Aging at the Albert Einstein College of Medicine.

**Author Contributions** X.D. and B.M. performed data analysis. X.D., B.M. and J.V. wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.V. ([jan.vijg@einstein.yu.edu](mailto:jan.vijg@einstein.yu.edu)).

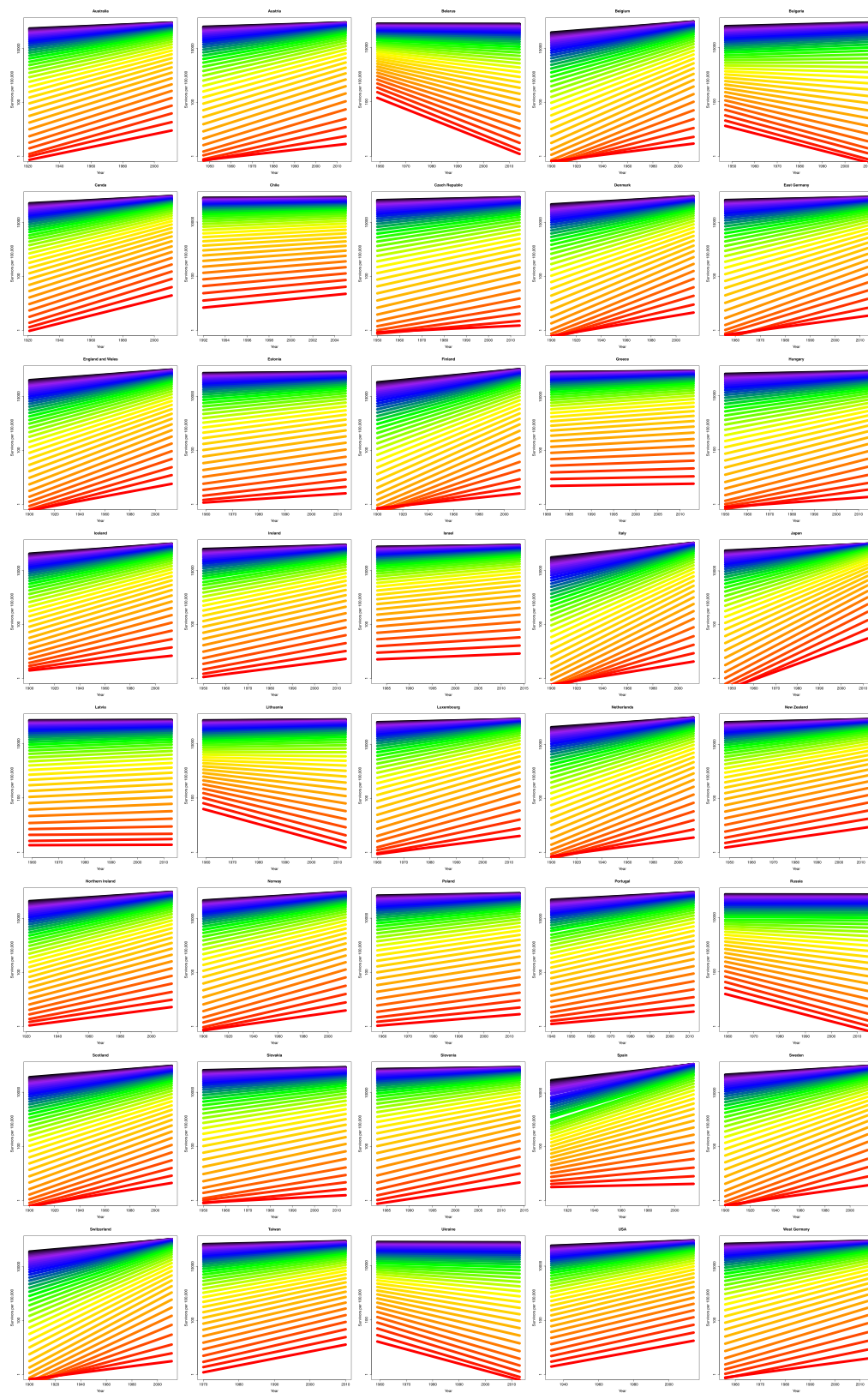
**Reviewer Information** *Nature* thanks J.-M. Robine and the other anonymous reviewer(s) for their contribution to the peer review of this work.



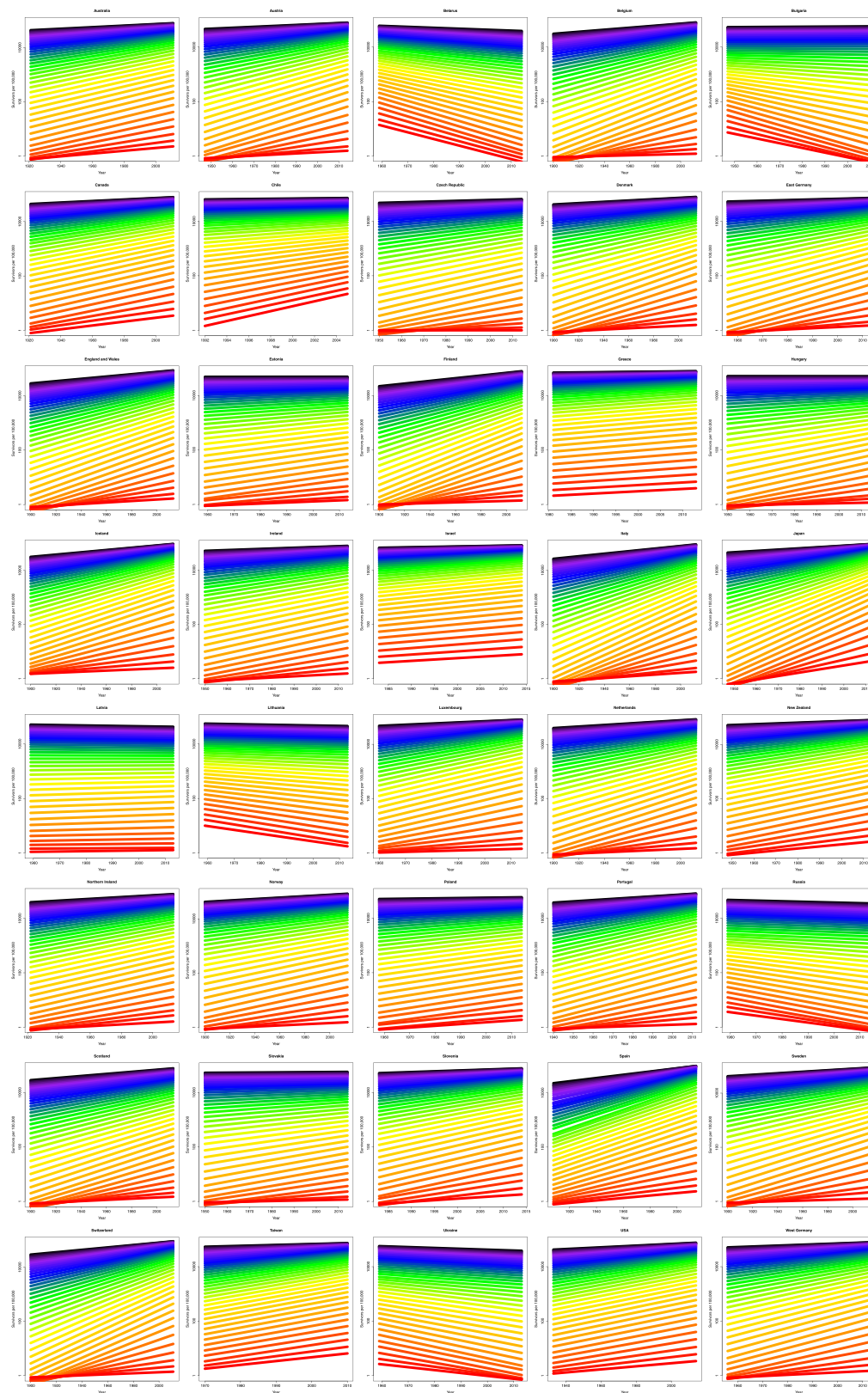


**Extended Data Figure 1 | Life expectancy over time since 1900 (or the earliest year for which data was available) in 40 countries and territories.** There is a generally positive trend over time; life expectancy in Japan appears to be reaching a plateau, but the increase looks unabated

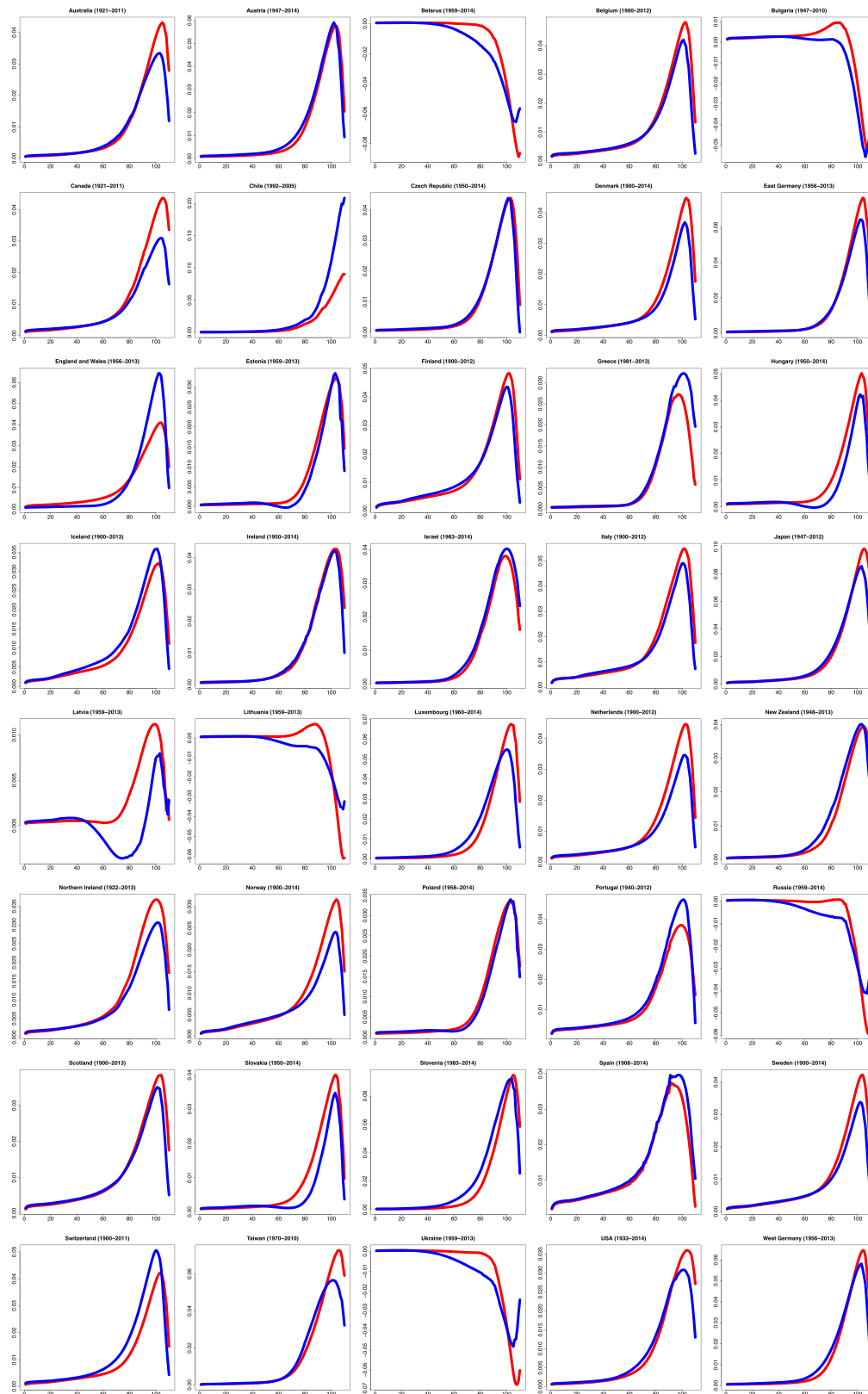
in many of the other countries. The data represent the entire population for each region, except Scotland, where it represents only the civilian population. The colour scheme is as in Fig. 1a.



**Extended Data Figure 2 | Proportion of the population surviving to old age among females in 40 countries and territories.** The data represent the entire population for each region, except Scotland, where it represents only the civilian population. The colour scheme is as in Fig. 1b.



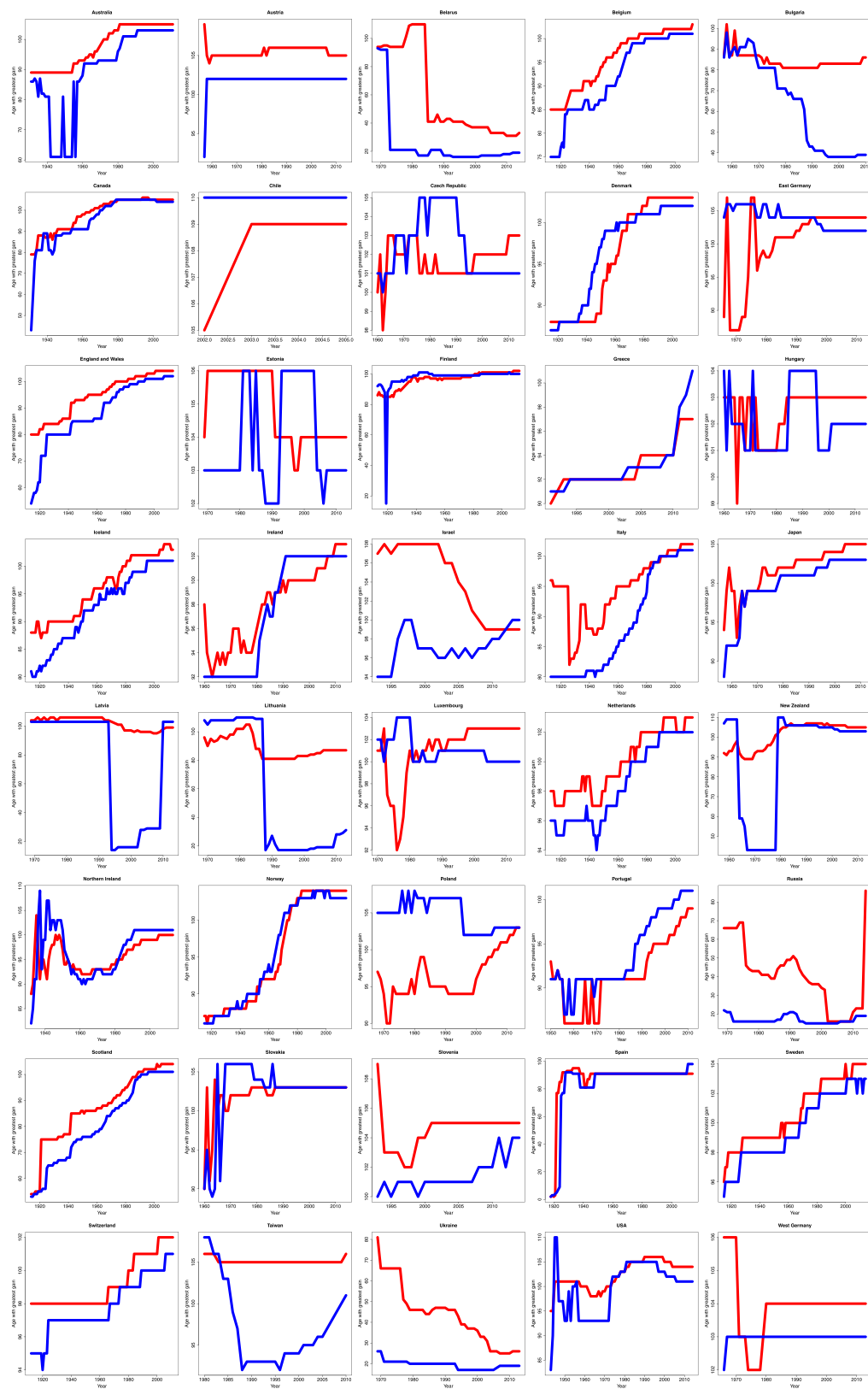
**Extended Data Figure 3 | Proportion of the population surviving to old age among males in 40 countries and territories.** The data represent the entire population for each region, except Scotland, where it represents only the civilian population. The colour scheme is as in Fig. 1b.



**Extended Data Figure 4 | Rate of change in survival since 1900 (or the earliest year for which data was available) to a given age as a function of that age in 40 countries and territories.** The rate of change is the slope of the line calculated by an exponential regression, that is,  $b$  in the

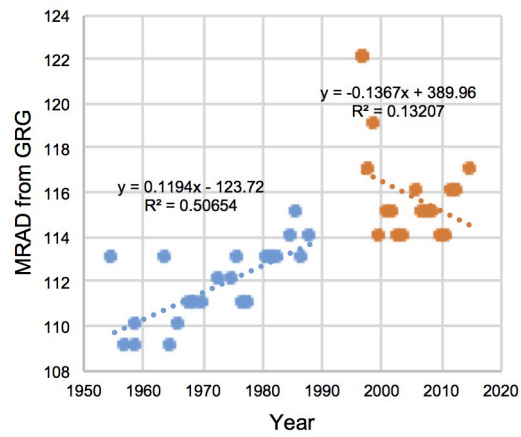
equation  $y = a + bx$ , where  $x$  is age and  $y$  is the logarithm of the number of survivors to that age per 100,000. Including France, 90% (37/41) of the regions examined exhibited the pattern depicted in Fig. 1c.





**Extended Data Figure 5 | Age with the greatest increase in survival as a function of calendar year in 40 countries and territories.** For each year, the age with the greatest increase in survival over the past 100 years (or since the earliest year for which data was available), that is, the peak of a graph like that from Extended Data Fig. 4, was determined. Including France (Fig. 1d), a total of 82 data sets were considered (males and females in each region); we used linear regressions of segments of the data to look for evidence of plateaus. A data set was considered to be plateauing if one

of the following criteria applied: the second half of the data had a negative slope; the first half of the data had a negative slope (as an increase in the second half would likely reflect a return to some equilibrium after being negatively perturbed); the first half of the data had a slope greater than that of the second half of the data; or the final 10% of the data had a slope less than that of the preceding 40%. In 88% (72/82) of the data sets, there was evidence of a plateau.



**Extended Data Figure 6 | The yearly maximum reported age at death from the GRG database (worldwide, 1972–2015).** The lines represent the functions of linear regressions.

# Tissue-specific mutation accumulation in human adult stem cells during life

Francis Blokzijl<sup>1,2</sup>, Joep de Lig<sup>1,2\*</sup>, Myrthe Jager<sup>1,2\*</sup>, Valentina Sasselli<sup>2\*</sup>, Sophie Roerink<sup>3\*</sup>, Nobuo Sasaki<sup>2</sup>, Meritxell Huch<sup>2</sup>, Sander Boymans<sup>1,2</sup>, Ewart Kuijk<sup>1,2</sup>, Pjotr Prins<sup>2</sup>, Isaac J. Nijman<sup>2</sup>, Inigo Martincorena<sup>3</sup>, Michal Mokry<sup>4</sup>, Caroline L. Wiegerinck<sup>4</sup>, Sabine Middendorp<sup>4</sup>, Toshiro Sato<sup>2</sup>, Gerald Schwank<sup>2</sup>, Edward E. S. Nieuwenhuis<sup>4</sup>, Monique M. A. Versteegen<sup>5</sup>, Luc J. W. van der Laan<sup>5</sup>, Jeroen de Jonge<sup>5</sup>, Jan N. M. IJzermans<sup>5</sup>, Robert G. Vries<sup>6</sup>, Marc van de Wetering<sup>2</sup>, Michael R. Stratton<sup>3</sup>, Hans Clevers<sup>2</sup>, Edwin Cuppen<sup>1,2</sup> & Ruben van Bostel<sup>1,2</sup>

**The gradual accumulation of genetic mutations in human adult stem cells (ASCs) during life is associated with various age-related diseases, including cancer<sup>1,2</sup>. Extreme variation in cancer risk across tissues was recently proposed to depend on the lifetime number of ASC divisions, owing to unavoidable random mutations that arise during DNA replication<sup>1</sup>. However, the rates and patterns of mutations in normal ASCs remain unknown. Here we determine genome-wide mutation patterns in ASCs of the small intestine, colon and liver of human donors with ages ranging from 3 to 87 years by sequencing clonal organoid cultures derived from primary multipotent cells<sup>3–5</sup>. Our results show that mutations accumulate steadily over time in all of the assessed tissue types, at a rate of approximately 40 novel mutations per year, despite the large variation in cancer incidence among these tissues<sup>1</sup>. Liver ASCs, however, have different mutation spectra compared to those of the colon and small intestine. Mutational signature analysis reveals that this difference can be attributed to spontaneous deamination of methylated cytosine residues in the colon and small intestine, probably reflecting their high ASC division rate. In liver, a signature with an as-yet-unknown underlying mechanism is predominant. Mutation spectra of driver genes in cancer show high similarity to the tissue-specific ASC mutation spectra, suggesting that intrinsic mutational processes in ASCs can initiate tumorigenesis. Notably, the inter-individual variation in mutation rate and spectra are low, suggesting tissue-specific activity of common mutational processes throughout life.**

It has not yet been possible to measure somatic mutation loads in ASCs from specific human tissues. However, such knowledge could be valuable in understanding tissue homeostasis and repair capacities as well as ASC vulnerabilities to extrinsic factors. The accumulation of mutations as life progresses is thought to underlie the genesis of age-related diseases such as cancer<sup>6</sup> and organ failure<sup>2</sup>. Mutations acquired in the genomes of multipotent ASCs are believed to have the largest impact on the mutational load of tissues, owing both to their potential for self-renewal and capacity to propagate mutations to their daughter cells<sup>1,2</sup>. Consistently, cancer-initiating mutations in intestinal ASCs lead to tumour formation within weeks, whereas these mutations fail to drive intestinal adenomas when induced in differentiated cells<sup>7</sup>. Unavoidable random mutations that arise during DNA replication in normal ASCs have recently been proposed to impart a large influence on cancer risk<sup>1</sup>. Consequently, tissues with a high ASC turnover would show higher cancer incidence when compared to tissues with low ASC proliferation rates<sup>1,8</sup>. However,

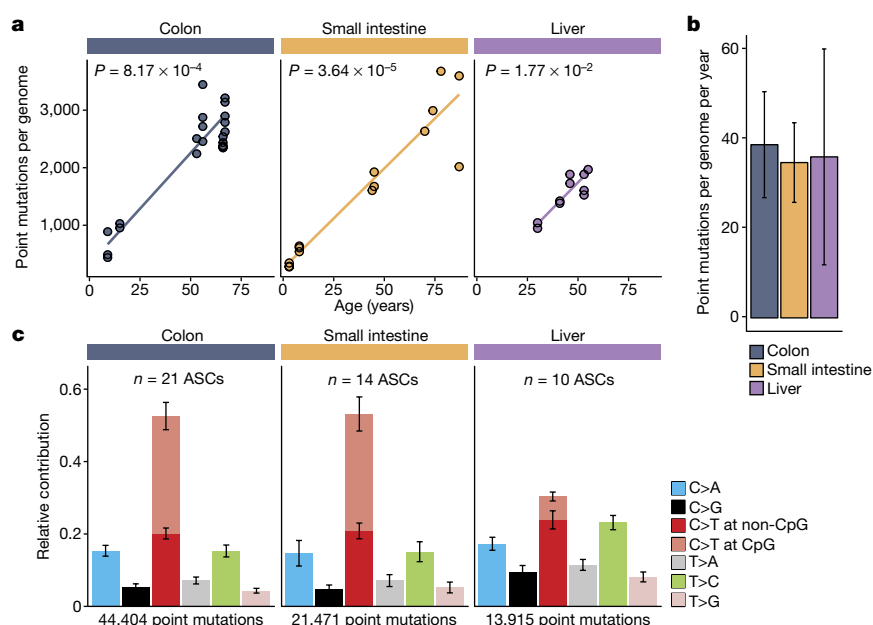
computational modelling has suggested that the variation in ASC proliferation rate alone cannot exclude extrinsic risk factors as important determinants of organ-specific cancer incidence<sup>9</sup>. Yet, the number of mutations that accumulate during the lifespan of normal human ASCs with different turnover rates has, to date, not been directly determined and compared. To understand tissue homeostasis and tissue-specific susceptibility to cancer and ageing-associated diseases it is important to assess mutation accumulation in ASCs of different tissues.

Here, we experimentally define ASCs as those cells that give rise to long-term organoid cultures and have the potential to differentiate into multiple tissue-specific cell types<sup>3–5</sup>. To catalogue the *in vivo*-acquired somatic mutations in individual normal human ASC genomes, we used an *in vitro* system to expand single ASCs into epithelial organoids, which reflect the genetic make-up of the original ASC (Extended Data Fig. 1a and Methods). This procedure allowed us to obtain sufficient DNA for accurate whole-genome sequencing (WGS) analysis, while circumventing the high noise levels associated with single-cell DNA amplification<sup>10</sup>. We assessed ASCs from the small intestine, colon and liver, tissues that differ greatly in proliferation rate and cancer risk<sup>1</sup>. Cancer incidence is much higher in the colon compared to the small intestine and liver<sup>1</sup>. We sequenced 45 independent clonal organoid cultures derived from 19 donors ranging in age from 3 to 87 years (Extended Data Table 1). In addition, we sequenced a blood or polyclonal biopsy sample of each donor to identify and exclude germline variants. Subclonal mutations, which must have been introduced *in vitro* after the single-cell step, were discarded based on their low variant-allele frequency (Extended Data Figs 1b–d, 2 and Methods). Overall, we identified 79,790 heterozygous clonal somatic point mutations and subsequent extensive validations showed an overall confirmation rate of approximately 91% (Extended Data Figs 1, 3).

A positive correlation (*t*-test linear mixed model;  $P < 0.05$ ) between the number of somatic point mutations and the age of the donor could be observed for all organs (Fig. 1a and Extended Data Fig. 4), indicating that ASCs gradually accumulate mutations with age, independent of tissue type. Notably, we found that the annual mutation rate in ASCs was in the same range for all assessed tissues, despite the dissimilar cancer incidence in these tissues; ASCs of the colon, small intestine and liver accumulate around 36 mutations per year (95% confidence intervals are 26.9–50.6, 25.8–43.6 and 11.9–60.1, respectively; Fig. 1b). The mutation spectra in small intestinal and colon ASCs were very similar, but differed markedly from liver (Fig. 1c). Notably, the mutation

<sup>1</sup>Center for Molecular Medicine, Cancer Genomics Netherlands, Department of Genetics, University Medical Center Utrecht, Heidelberglaan 100, 3584CX Utrecht, The Netherlands. <sup>2</sup>Hubrecht Institute for Developmental Biology and Stem Cell Research, KNAW and University Medical Center Utrecht, Uppsalalaan 8, 3584CT Utrecht, The Netherlands. <sup>3</sup>Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. <sup>4</sup>Department of Pediatrics, University Medical Center Utrecht, Lundlaan 6, 3584 EA Utrecht, The Netherlands. <sup>5</sup>Department of Surgery, Erasmus MC-University Medical Center, Postbus 2040, 3000 CA Rotterdam, The Netherlands. <sup>6</sup>Foundation Hubrecht Organoid Technology (HUB), Uppsalalaan 8, 3584CT Utrecht, The Netherlands.

\*These authors contributed equally to this work.



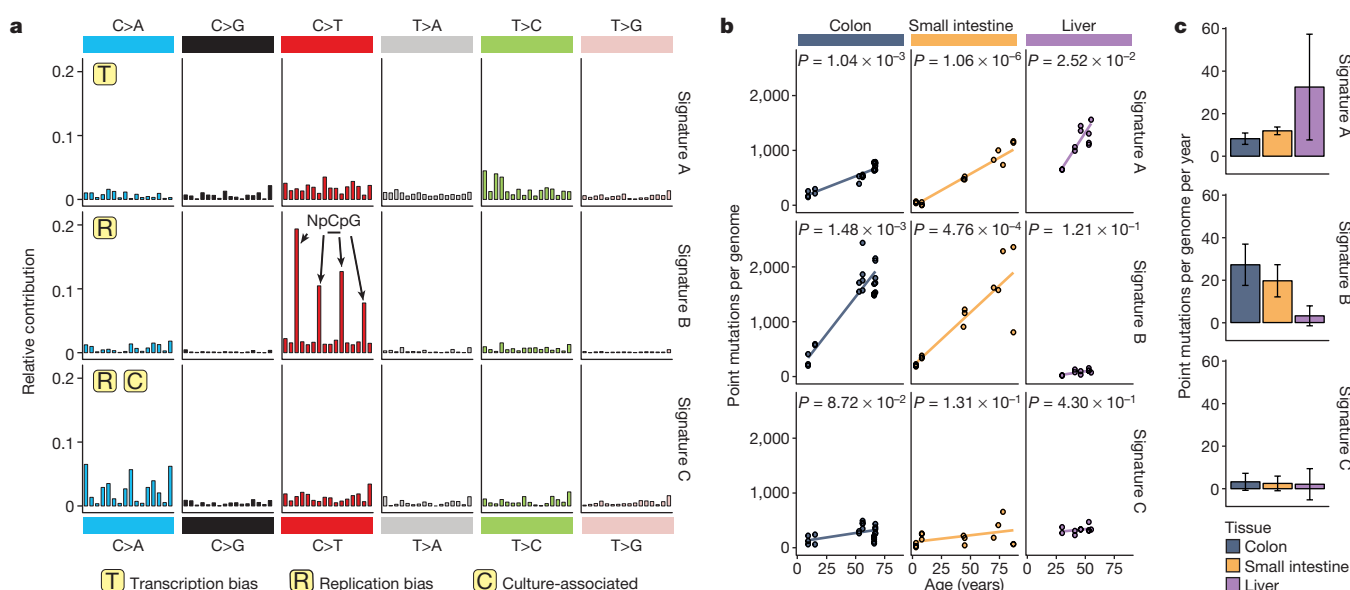
**Figure 1 | Age-associated accumulation of somatic point mutations in human ASCs.** **a**, Correlation of the number of somatic point mutations in each ASC type examined (extrapolated to the whole autosomal genome) with age of the donors per tissue. Each data point represents a single ASC. The  $P$  values of the age effects in the linear mixed model (two-tailed  $t$ -test) are indicated for each tissue. The sample sizes for colon, small intestine and liver ASCs are 6, 9 and, 5 donors, with, in total, 21, 14 and 10 ASCs, respectively. **b**, Somatic mutation accumulation rate per tissue as estimated by the linear mixed models in **a**. Error bars represent the 95% confidence intervals of the slope estimates. **c**, Relative contribution of the indicated mutation types to the point mutation spectrum for each tissue type. Data are represented as the mean relative contribution of each mutation type over all ASCs per tissue type ( $n = 21, 14$  and  $10$  for colon, small intestine and liver, respectively) and error bars represent standard deviation. The total number of identified somatic point mutations per tissue is indicated.

spectrum within tissues did not differ between young and elderly donors (Extended Data Fig. 5).

Genome-wide mutation patterns in the ASCs provide insights into the mutational and DNA repair processes that are active in different organs<sup>11</sup>. Using non-negative matrix factorization<sup>12</sup>, we extracted three mutational process signatures (Fig. 2a and Methods). All of these signatures were previously described in a pan-cancer analysis<sup>11</sup>. Signature A (corresponding to signature 5 in ref. 11), characterized by T:A to C:G transitions, was the main contributor to the mutation spectrum observed in the liver and was also clearly present in the small intestine and colon (Fig. 2). Although the underlying mutational process remains unknown, the number of mutations attributed

to this signature that accumulate with age resembles a linear trend in all tissues (Fig. 2b). This suggests that this signature represents a universal genomic ageing mechanism (that is, a chemical process acting on DNA molecules) independent of cellular function or proliferation rate.

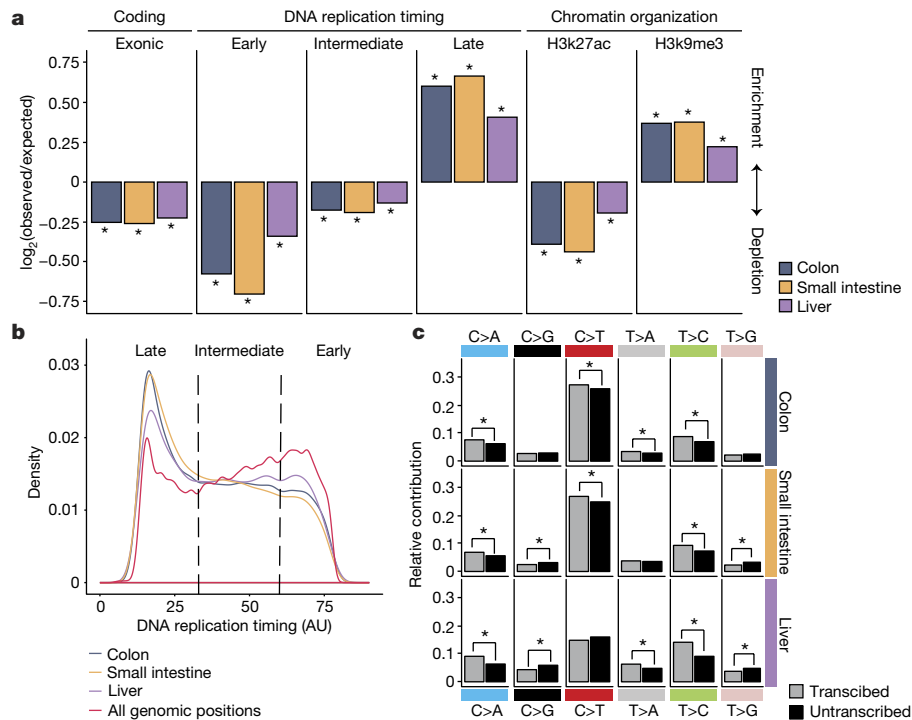
The majority of the somatic mutations observed in small intestinal and colon ASCs could be attributed to signature B (corresponding to signature 1A in ref. 11), which is characteristic of spontaneous deamination of methylated cytosine residues into thymine at CpG sites (Fig. 2a). The resulting T:G mismatch can be effectively repaired, but the mutation is incorporated if DNA replication occurs before the repair is initiated<sup>13</sup>. In line with this, high rates of signature B mutations



**Figure 2 | Signatures of mutational processes in human ASCs and their tissue-specific contribution.** **a**, Contribution of context-dependent mutation types to the three mutational signatures that were identified by non-negative matrix factorization (NMF) analysis of the somatic mutation collection observed in the ASCs across all assessed tissues. The contribution of each trinucleotide (order is similar to that in ref. 11) to each signature is shown. For each signature, the presence of transcriptional-strand bias, DNA-replication-timing bias and/or association with the culture system

is indicated. **b**, Absolute contribution of each mutational signature type (extrapolated to the whole autosomal genome) plotted against the age of the donors for each tissue. Each data point represents a single ASC. The  $P$  values of the age effects per tissue are shown (linear mixed model, two-tailed  $t$ -test). **c**, Signature-specific mutation rate per year per genome for each tissue as estimated by the linear mixed model in **b**. Error bars represent the 95% confidence intervals of the slope estimates.





**Figure 3 | Non-random genomic distribution of somatic point mutations in ASCs.** **a**, Enrichment and depletion of somatic point mutations in the indicated genomic regions for each tissue. The  $\log_2$  ratio of the number of observed and expected point mutations indicates the effect size of the enrichment or depletion in each region.  $*P < 0.05$ ,

one-sided binomial test. **b**, Distribution of DNA replication timing for all genomic positions and the somatic point mutations detected in human ASCs per tissue. **c**, Relative contribution of each point-mutation type on the transcribed and untranscribed strand for each tissue.  $*P < 0.05$ , two-sided Poisson test.

are observed in many cancer types of epithelial origin with high cell turnover<sup>13</sup>. This process showed a minimal contribution to the age-related mutational load in liver ASCs (Fig. 2c), which is likely to reflect the relatively low division rate of these cells during life. Finally, contribution of a third signature, signature C (corresponding to signature 18 in ref. 11), was minimal in all tissues and did not correlate with age (Fig. 2b). Sequential clonal ASC expansions in culture followed by WGS analysis showed that *in vitro*-induced mutations are predominantly characterized by this signature (Extended Data Fig. 6 and Methods).

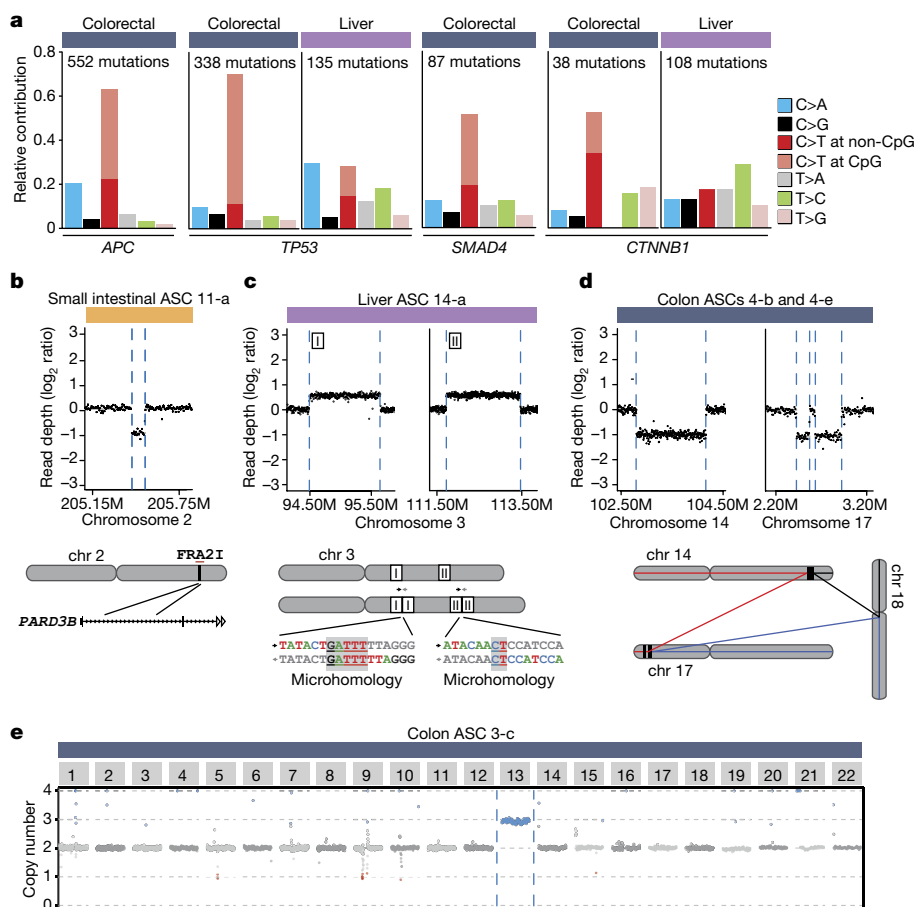
Signature B mutations were strongly associated with the timing of replication and predominantly present in late-replicating DNA (Extended Data Fig. 7) even though the majority of CpG dinucleotides are located in early-replicating DNA. This bias suggests that this mutagenic process is more active in late-replicating DNA or, alternatively, that replication-coupled repair shows reduced activity in late-replicating DNA<sup>14</sup>. Consequently, somatic mutations in small intestine and colon ASCs were strongly enriched in late-replicating DNA and depleted in early-replicating DNA (Fig. 3a, b). In addition, somatic point mutations in small intestine and colon ASCs were depleted in H3K27ac (histone H3 acetyl Lys27)-associated DNA and enriched in H3K9me3 (histone H3 trimethyl Lys9)-associated DNA (Fig. 3a), similar to patterns previously observed in cancer<sup>15</sup>. As genic regions are predominantly located in early-replicating DNA and open chromatin, we observed a depletion of mutations in exonic sequences (Fig. 3a). This demonstrates that genome-wide mutation rates and spectra cannot be reliably estimated using mutation discovery in reporter genes<sup>16</sup>, such as the T-lymphocyte *HPRT* cloning assay<sup>17</sup>, or by deep sequencing of genic regions<sup>18–21</sup>. To test whether the depletion of coding mutations was caused by selection against cells with damaging mutations, we calculated the ratio of non-synonymous to synonymous mutations (dN/dS) taking into account the mutation spectra and sequence composition (see Methods)<sup>18</sup>. We did not observe negative selection for non-synonymous mutations

(Extended Data Fig. 7f), arguing against the negative selection of cells with damaging protein-coding mutations.

In liver ASCs, somatic mutations are more randomly distributed throughout the genome and are less associated with replication timing or chromatin status (Fig. 3a). Nevertheless, a comparable depletion of exonic mutations was observed in all tissues (Fig. 3a), suggesting that liver ASCs use different mechanisms to maintain genetic integrity in functionally relevant regions. Signature A, the most predominant in liver ASCs, shows little bias towards DNA-replication-timing dynamics, but a pronounced transcriptional-strand bias<sup>11</sup> (Extended Data Fig. 7), consistent with activity of transcription-coupled repair<sup>22</sup>. In line with this, point mutations in the genic regions of the assessed ASCs showed a significant transcriptional strand bias, exemplified by the more frequent occurrence of T:A to C:G transitions on the transcribed strand compared to the untranscribed strand (Fig. 3c).

Our results indicate that a stable balance between the degree of DNA damage and the subsequent repair is maintained throughout life in various ASC types, since mutations accumulate steadily and display a constant mutation spectrum. Earlier work in mice using mutation-discovery in a *LacZ* reporter gene, showed major age-related changes in mutation spectra in different tissues<sup>23</sup>. The difference between these observations could be explained by the comprehensive genome-wide analysis applied here to ASCs, whereas reporter assays assess specific genes predominantly in differentiated cells. Although variation in tissue-specific mutation spectra in mice has been reported previously<sup>23–25</sup>, we observed a difference in both mutation rate and spectrum in human cells (Extended Data Fig. 8). This indicates that mutation data derived from mice are not necessarily suitable for interpreting mutational processes and their consequences in humans.

Although we analysed cells from many different donors without controlling for lifestyle differences or gender, the point-mutation rate and spectrum were highly similar between individuals within organs. This suggests that incidental exposure to environmental mutagenic factors has minimal effect on the point-mutation landscapes in



**Figure 4 | Cancer-associated mutation spectra in driver genes and structural variation in normal ASCs.** **a**, Spectrum of point mutations in cancer driver genes *APC*, *TP53*, *SMAD4* and *CTNNB1* identified in colorectal and liver cancer. The total number of somatic point mutations per gene per cancer type is indicated. **b**, Read-depth analysis indicating a relatively small deletion (~90 kb) located within a common fragile site (*FRA2I*) in intestinal ASC 11-a. Each point represents the  $\log_2$  value of the GC-corrected read-depth ratio per 5-kb window. Dashed lines indicate breakpoint regions; a schematic representation of the identified structural variant with associated genomic and breakpoint features is depicted below.

**c**, Two large (>1 Mb) tandem duplications identified in liver ASC 14-a with microhomology at the breakpoints; duplications are indicated in the schematic representation of the identified structural variants below the graph. **d**, A complex structural variation (an unbalanced translocation involving 3 chromosomes) identified in colon ASCs 4-b and 4-e. Coloured lines in the schematic below show the predicted derivative chromosomes. **e**, Read-depth analysis indicating a trisomy of chromosome 13 in colon ASC 3-c. Each data point represents the median chromosome copy number per 500-kb bin plotted over the genome, with alternating colours for each successive chromosome.

normal ASCs of the organs we assessed. Cell-intrinsic mutational processes, such as deamination-induced mutagenesis in rapidly cycling ASCs, seem to be more important determinants of point-mutation load. Indeed, many colorectal cancer mutations in the driver genes *APC*, *TP53*, *SMAD4* and *CTNNB1* are C:G to T:A transitions at CpG dinucleotides, whereas liver cancer driver mutations in the same genes have a completely different spectrum (Fig. 4a). However, ASCs of the colon and small intestine show very similar age-related mutation characteristics, although cancer incidence is extremely low in the human small intestine<sup>1,9</sup>. In addition to somatic point mutations, we evaluated the presence of somatic structural variants (Fig. 4b–e and Extended Data Table 2). We detected small deletions (91–443 kb) in 3 out of 14 small intestinal ASCs and a larger deletion (2 Mb) in one ASC. Notably, colon ASCs showed complex and larger chromosomal instability in 4 out of 15 colon ASCs, including a complex translocation (Fig. 4d) and a trisomy (Fig. 4e). These events are characteristic of segregation errors that can occur during cell division, and are a hallmark of many colorectal cancers<sup>26</sup>. In addition, other factors, such as tissue clonality or external agents may also contribute to the difference in cancer incidence between colon and small intestine.

Here we have shown that ASCs of organs with different cancer incidences gradually accumulate mutations at similar rates, but that the

mutation profiles are tissue-specific. In the ASCs of the tissues assessed here, mutation accumulation is primarily driven by a combination of proliferation-dependent mutation incorporation following spontaneous deamination of methylated cytosine residues and another process with a currently unknown underlying molecular mechanism. Notably, the former intrinsic, unavoidable mutational process can cause the same types of mutation as those observed in cancer driver genes. We have shown that, at least in colon ASCs, this class of mutations could have a role in driving tumorigenesis.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 7 March; accepted 16 August 2016.**

**Published online 3 October 2016.**

1. Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
2. Rossi, D. J., Jamieson, C. H. M. & Weissman, I. L. Stems cells and the pathways to aging and cancer. *Cell* **132**, 681–696 (2008).
3. Huch, M. *et al.* Long-term culture of genome-stable bipotent stem cells from adult human liver. *Cell* **160**, 299–312 (2015).
4. Sato, T. *et al.* Single Lgr5 stem cells build crypt-villus structures *in vitro* without a mesenchymal niche. *Nature* **459**, 262–265 (2009).

5. Sato, T. *et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).
6. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
7. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608–611 (2009).
8. Milholland, B., Auton, A., Suh, Y. & Vijg, J. Age-related somatic mutations in the cancer genome. *Oncotarget* **6**, 24627–24635 (2015).
9. Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–47 (2016).
10. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
11. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
12. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3**, 246–259 (2013).
13. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
14. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
15. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
16. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).
17. Finette, B. A. *et al.* Determination of *HPRT* mutant frequencies in T-lymphocytes from a healthy pediatric population: statistical comparison between newborn, children and adult mutant frequencies, cloning efficiency and age. *Mutat. Res.* **308**, 223–231 (1994).
18. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
19. Xie, M. *et al.* Age-related cancer mutations associated with clonal hematopoietic expansion. *Nat. Med.* **20**, 1472–1478 (2014).
20. Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
21. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
22. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
23. Dollé, M. E. T., Snyder, W. K., Dunson, D. B. & Vijg, J. Mutational fingerprints of aging. *Nucleic Acids Res.* **30**, 545–549 (2002).
24. Dollé, M. E., Snyder, W. K., Gossen, J. A., Lohman, P. H. & Vijg, J. Distinct spectra of somatic mutations accumulated with age in mouse heart and small intestine. *Proc. Natl Acad. Sci. USA* **97**, 8403–8408 (2000).
25. Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
26. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **6**, 479–507 (2011).

**Acknowledgements** The authors would like to thank the gastroenterologists of the UMCU/Wilhelmina Children's Hospital and Diaconessen Hospital for obtaining human duodenal and colon biopsies and R. Eijkemans for his advice on the statistical analyses. This study was financially supported by a Zenith grant of the Netherlands Genomics Initiative (935.12.003) to E.C., the NWO Zwaartekracht program Cancer Genomics.nl and funding of Worldwide Cancer Research (WCR no. 16-0193) to R.B. We declare no competing financial interests.

**Author Contributions** C.L.W., S.M. and E.E.S.N. obtained duodenal biopsies. N.S., M.M., E.E.S.N., M.M.A.V. and J.J. obtained colon biopsies. M.M.A.V., L.J.W.L., J.J. and J.N.M.I. obtained human liver biopsies. M.J., V.S., N.S., M.H., E.K., C.L.W., T.S., G.S. and R.B. performed ASC culturing. M.W. performed cell sorting. S.R., M.R.S., E.C. and R.B. performed sequencing. F.B., J.L., S.B., P.P., I.J.N., I.M. and R.B. performed bioinformatic analyses. F.B., R.G.V., H.C., E.C. and R.B. were involved in the conceptual design of the study. F.B., H.C., E.C. and R.B. wrote the manuscript.

**Author Information** The human sequencing data have been deposited at the European Genome-phenome Archive (<http://www.ebi.ac.uk/ega/>) under accession numbers EGAS00001001682 and EGAS00001000881. The mouse sequencing data have been deposited at the European Nucleotide Archive (<http://www.ebi.ac.uk/ena/>) under accession number ERP005717. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.C. ([ecuppen@umcutrecht.nl](mailto:ecuppen@umcutrecht.nl)).

**Reviewer Information** *Nature* thanks G. Pfeifer, L. Vermeulen, J. Vijg and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

No sample-size estimate was calculated before the study was executed. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Human tissue material.** Endoscopic, colorectal and duodenal biopsy samples were obtained from individuals of different ages that had been admitted for suspected inflammation. One individual (donor 1) showed no inflammation during colonoscopy, but was later diagnosed with microscopic colitis. The other individuals were found to be healthy based on standard histological examination. Endoscopic biopsies were performed at the University Medical Center Utrecht and the Wilhelmina Children's Hospital. The patients' informed consent was obtained and this study was approved by the ethical committee of University Medical Center Utrecht. Additionally, normal tissue was isolated from resected colon segments at >5 cm distance from a tumour in three colorectal cancer patients (donors 3, 4 and 19). The colonic tissues were obtained at The Diaconessen Hospital Utrecht with informed consent and the study was approved by the ethical committee. Liver biopsies (0.5–1 cm<sup>3</sup>) were obtained from donor livers during transplantations performed at the Erasmus Medical Center, Rotterdam. Both liver and colon biopsies were obtained from donor 18. The Medical Ethical Council of the Erasmus MC approved the use of this material for research purposes, and informed consent was provided by all donors and/or relatives.

**Establishment of clonal ASC cultures.** Dissociated colon and small intestinal crypts were isolated from the biopsies and cultured for 1–2 weeks under conditions that are optimal for stem-cell proliferation, as previously described<sup>5</sup>. Liver cells were isolated from human liver biopsies and cultured as previously described<sup>3</sup>. From these cultures, single cells were sorted by flow cytometry and clonally expanded (Extended Data Fig. 1a). Clonal ASC cultures were subsequently established by manual picking of individual organoids derived from single cells and *in vitro* expansion for a period of ~6 weeks.

**Whole-genome sequencing and read alignment.** DNA libraries for Illumina sequencing were generated using standard protocols (Illumina) from 200 ng–1 µg of genomic DNA isolated from the clonally expanded ASC cultures with genomic tips (Qiagen). The libraries were sequenced with paired-end (2 × 100 bp) runs using Illumina HiSeq 2500 sequencers to a minimal depth of 30× base coverage. Samples of donors 1, 2, 3, 4, 10, 12, 13, 15, 16, 18 and 19 were sequenced using Illumina HiSeq X Ten sequencers to equal depth. The reference samples, blood or biopsy, were sequenced similarly. Sequence reads were mapped against human reference genome GRCh37 using Burrows–Wheeler Aligner v0.5.9 mapping tool<sup>27</sup> with settings 'bwa mem -c 100 -M'. Sequence reads were marked for duplicates using Sambamba v0.4.7 (ref. 28) and realigned per donor using Genome Analysis Toolkit (GATK) IndelRealigner v2.7.2 and sequence read-quality scores were recalibrated with GATK BaseRecalibrator v2.7.2. Alignments from different libraries of the same ASC culture were combined into a single BAM file.

**Point mutation calling.** Raw variants were multi-sample (per donor) called using the GATK UnifiedGenotyper v2.7.2 (ref. 29) and GATK-Queue v2.7.2 with default settings and additional option 'EMIT\_ALL\_CONFIDENT\_SITES'. The quality of variant and reference positions was evaluated using GATK VariantFiltration v2.7.2 with options '-filterExpression "MQ0 ≥ 4 && (MQ0 / (1.0 \* DP)) > 0.1"'-filterName "HARD\_TO\_VALIDATE"-filterExpression "QUAL < 30.0"-filterName "VeryLowQual"-filterExpression "QUAL > 30.0 && QUAL < 50.0"-filterName "LowQual"-filterExpression "QD < 1.5"-filterName "LowQD".

**Point mutation filtering.** To obtain high-quality catalogues of somatic point mutations, we applied a comprehensive filtering procedure (Extended Data Fig. 1b). We considered variants that were passed by VariantFiltration and had a GATK phred-scaled quality score ≥ 100. Subsequently, for each ASC culture, we considered the positions with a base coverage of at least 20× in both the culture and the reference sample (blood or biopsy). Furthermore, we only regarded variants at autosomal chromosomes. We excluded variant positions that overlapped with single-nucleotide polymorphisms (SNPs) in the SNP database (dbSNP) v137.b37 (ref. 30). Furthermore, we excluded all positions that were found to be variable in at least two of three unrelated individuals (that is, donor 5, 6 and X (not in study)) to exclude recurrent sequencing artefacts. To obtain somatic point mutations, we filtered out all variants with any evidence of the alternative allele in the reference sample. We validated the clonal origin of the sequenced ASC cultures by analysing the variant allele frequencies (VAFs) of the somatic mutations. Two cultures (donor 14, cell b and donor 17, cell c) showed a shift in the peak of the somatic heterozygous mutations to the left, indicating that they did not arise from a single stem cell, and were therefore excluded from the analysis (Extended Data Fig. 2). Finally, for all cultures we excluded point mutations with a VAF < 0.3 to exclude mutations that were potentially induced *in vitro* after the (first) clonal step (Extended Data Fig. 1b–d). The number of mutations that passed each filtering step for the samples of donor 5 and 6 is depicted in Extended Data Fig. 1c.

The overlap of the point mutations between ASCs of the same donor is depicted in Extended Data Fig. 4d.

**Validations of point mutations.** We evaluated our mutation filtering procedure by independent validations of 374 pre-selected positions that were either discarded or passed during filtering using amplicon-based next-generation sequencing. To this end, primers were designed ~250 nucleotides 5' and 3' from the candidate point mutations to obtain amplicons of ~500 bp (primer sequences available upon request). These regions were PCR-amplified for both the organoid cultures and reference samples of donor 5 and 6, using 5 ng genomic DNA, 1× PCR Gold Buffer (Life Technologies), 1.5 mM MgCl<sub>2</sub>, 0.2 mM of each dNTP and 1 unit of AmpliTaq Gold (Life Technologies) in a final volume of 10 µl. This which was held at 94°C for 60 s followed by 15 cycles at 92°C for 30 s, 65°C for 30 s (with a decrement of 0.2°C per cycle) and 72°C for 60 s; followed by 30 cycles of 92°C for 30 s, 58°C for 30 s and 72°C for 60 s; with a final extension at 72°C for 180 s. The PCR products were pooled and barcoded per culture. Illumina sequence libraries were generated according to the manufacturer's protocol. Subsequently, the libraries were pooled and sequenced using the MiSeq platform (2 × 250 bp) to an average depth of ~100×. Alignment and variant-calling was performed as described above. For each ASC we evaluated those positions with at least 20× coverage for both culture and reference sample, and defined positive positions as those with a call in culture, with a VAF ≥ 0.3 and no call in the reference sample. Subsequently, we determined the number of confirmed negatives of the positions that were filtered out for each filter step (Extended Data Fig. 1d). Moreover, we determined the number of confirmed positive of the positions that passed all filters (Extended Data Fig. 1e, f).

**Assessment of effects of *in vitro* culturing on ASC mutation load.** We expanded 10 initial clonal organoid cultures from small intestine and liver for a further 3–5 months (equivalent to ~20 weekly passages), upon which we isolated single cells and subjected them to clonal expansion to obtain sufficient DNA for WGS (Extended Data Fig. 6a). This approach allowed us to catalogue the mutations that accumulated in single ASCs during the culturing period between the two clonal steps. To this end, we selected the somatic point mutations that were unique to the sub-clonal cultures and not present in the corresponding original clonal cultures and therefore acquired during the *in vitro* expansion. We evaluated the specificity of our mutation-discovery procedure by determining the confirmation rate of the mutations identified in the original clone in the corresponding subclone. Only positions that had a coverage of ≥ 20× in both the original clonal and corresponding subclonal culture as well as in the reference sample were evaluated. On average, 91.1% ± 4.87 (mean ± s.d.) of these point mutations were confirmed in the subclonal cultures (Extended Data Fig. 3).

**Correlation between ASC somatic point mutation accumulation and age.** The surveyed area per ASC was calculated as the number of positions coverage ≥ 20× in both culture and the reference sample. The percentage of the whole non-N autosomal genome (GCRh37: 2,682,655,440 bp) that is surveyed in each ACS is depicted in Extended Data Table 1. For each ASC the total number of identified somatic point mutations was extrapolated to the whole non-N autosomal genome using its surveyed area. Subsequently, a linear mixed-effects regression model was fitted to estimate the effect of age on the number of somatic point mutations for each tissue using the nlme R package<sup>31,32</sup>, in which 'donor' is modelled as a random effect to resolve the non-independence that results from having multiple measurements per donor. A two-tailed *t*-test was performed to test whether the slope is significantly different from zero (that is to say, whether the fixed age effect in the linear mixed model is statistically significant). The intercept of the regression lines with the *y* axis represents the somatic mutations present at birth (that have accumulated in the tissue lineage during prenatal development) plus the noise levels in the data and the mutations that have accumulated during the first week(s) of culturing proceeding the clonal step (see above). Since all cells were assessed in a similar manner, noise levels will be comparable and therefore will not bias the mutation rate (slope) estimates. The slope of the regression line was used to estimate the fixed age effect on somatic point mutation rate per tissue.

To exclude the possibility that differences in surveyed areas between ASCs bias our results, we performed the age correlation and spectrum analyses on a subset of mutations that are located in genomic regions that are surveyed (≥ 20×) in all samples in this study. This consensus surveyed area comprises 38.2% of the autosomal non-N genome and both the mutation rate and spectra were highly similar to those in Fig. 1c (Extended Data Fig. 4a–c), indicating that the differences in surveyed areas between the clones do not bias our conclusions.

**Definition of genomic regions.** To generate a conserved DNA replication timing profile for the human genome, we downloaded 16 Repli-seq data sets from the ENCODE project<sup>33</sup> at the University of California, Santa Cruz (UCSC) genome browser<sup>34</sup> (GRCh37/hg19). The data consisted of Wavelet-smoothed values per 1-kb bin throughout the genome for 15 different cell lines (BJ, BG02ES, GM06990,



GM12801, GM12812, GM12813, GM12878, HeLa-S3, HepG2, HUVEC, IMR90, K562, MCF-7, NHEK and SK-N-SH). We considered the median values of all cell lines per bin, thereby excluding cell-specific values. We arbitrarily divided the genome into early- ( $\geq 60$ ), intermediate- ( $>33$  &  $<60$ ) and late- ( $\leq 33$ ) replicating bins (Fig. 3b). To generate a conserved chromatin-association profile for the human genome, we downloaded data containing the H3K9me3 signal per 25-nucleotide bin throughout the genome for 22 different cell lines (A549, AG04450, DND41, GM12878, H1-hESC, HeLa-S3, HepG2, HMEC, HSMM, HSMMt, HUVEC, K562, monocytes-CD14+\_RO1746, NH-A, NHDF-Ad, NHEK, NHLF, osteoblasts, MCF-7, NT2-D1, PBMC and U2OS) and the H3K27ac signal for 9 different cell lines (CD20+\_RO01794, DND41, H1-hESC, HeLa-S3, HSMM, monocytes-CD14+\_RO1746, NH-A, NHDF and osteoblasts). Data were downloaded from the ENCODE project<sup>33</sup> at the UCSC browser<sup>34</sup> (GCRh37/hg19) and the median values of all cell lines per bin were calculated. Next, we determined the distribution of the fractions of all bins (genome-wide). According to the shape of the resulting graph, we considered bins with an H3K9me3 value  $\geq 4$ , or an H3K27ac value  $\geq 2$ , as associated with that chromatin mark. Finally, exonic sequences were defined as all exonic regions reported in Ensembl v75 (GCRh37)<sup>35</sup>.

**Enrichment or depletion of point mutations in genomic regions.** We determined whether somatic point mutations were enriched or depleted in the genomic regions described above. To this end, we determined how many point mutations were observed in each genomic region for each donor. Next, we calculated the number of bases that were surveyed in each genomic region and calculated the expected number of point mutations by multiplying this surveyed length with the genome-wide point-mutation frequency. The  $\log_2(\text{observed/expected})$  of the mutations in the genomic regions was used as a measure of the effect size of the depletion or enrichment. One-tailed binomial tests were performed to calculate the statistical significance of deviations from the expected number of mutations in the genomic regions using  $\text{pbinom}$ <sup>31</sup>;  $P < 0.05$  was considered significant.

**Mutational signatures.** The occurrences of all 96-trinucleotide changes were counted for each ASC and averaged per donor. Three mutational signatures were extracted using NMF<sup>36</sup>. To determine the replication bias of signatures, we determined whether the point mutations were located in an intermediate, early or late replicating region (as defined above) using GenomicRanges<sup>37</sup> and repeated the NMF on a 288 count matrix (96 trinucleotides  $\times$  3 replication timing regions). Similarly, we looked at transcriptional strand bias by performing NMF on a 192 count matrix (96 trinucleotides  $\times$  2 strands). To this end, we selected all point mutations that fall within gene bodies and checked whether the mutated C or T was located on the transcribed or non-transcribed strand. We defined the transcribed units of all protein coding genes based on Ensembl v75 (GCRh37)<sup>35</sup> and included introns and untranslated regions.

**Selection analysis (dN/dS).** The dN/dS ratio was determined as described previously<sup>18</sup>. In brief, we used 192 rates, one for each of the possible trinucleotide changes in both strands. For each substitution type, we counted the number of potential synonymous and non-synonymous mutations in the protein-coding sequences of the human genome, using the longest DNA coding sequence as the reference sequence for each gene. Poisson regression was used to obtain maximum-likelihood estimates and confidence intervals of the normalized ratio of non-synonymous versus synonymous mutations (dN/dS ratio). The dN/dS ratio was tested against neutrality (dN/dS = 1) using a likelihood-ratio test.

**Comparison of mouse and human intestinal ASCs mutation loads.** Intestinal ASCs were isolated from the proximal part of the small intestine of randomly chosen ~2-year-old mice (one male and one female) carrying the Lgr5-EGFP-Ires-CreERT2 allele (mice were C57BL/6 background) by sorting for GFP<sup>high</sup> cells. Subsequently, three Lgr5-positive cells per animal were clonally expanded as described<sup>4</sup>. All experiments were approved by the Animal Care Committee of the Royal Dutch Academy of Sciences according to the Dutch legal ethical guidelines. DNA isolated from the intestinal ASC cultures isolated from mouse 1 were sequenced with paired-end (75 and 35 bp) runs using SOLiD 5500 sequencers (Life Technologies) to an average depth of  $\sim 18\times$  base coverage. Intestinal ASC cultures of mouse 2 were sequenced using Illumina HiSeq 2500 sequencers as described above. Sequence reads were aligned using Burrows–Wheeler Aligner to the mouse reference genome (NCBIM37) and point mutations were called using

the GATK UnifiedGenotyper v2.7.2 as described above. Post-processing filters for the intestinal ASCs of mouse 1 (analysed by SOLiD sequencing) were as follows: a minimum depth of  $10\times$ , variant uniquely called in one intestinal stem cell without more than one alternative allele found at the same position in the other ASCs of the same mouse, a GATK a phred-scaled quality score  $\geq 100$ , variant absent in mouse 2, variant position absent in the dbSNP (build 128) and a VAF  $\geq 0.25$ . Post-processing filters for the intestinal ASCs of mouse 2 (analysed by Illumina sequencing) were as described above for the human mutation data.

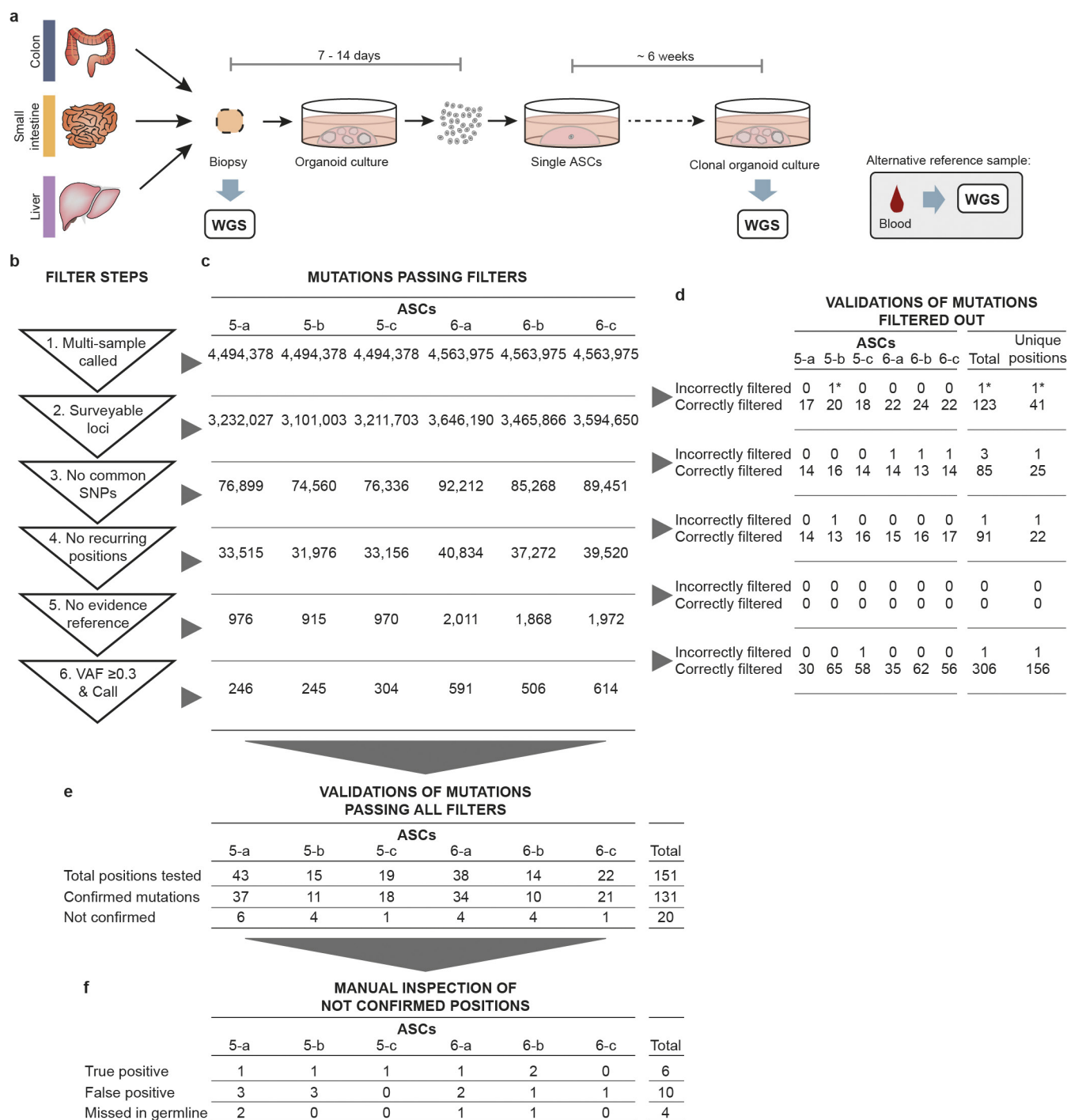
**Cancer-associated mutation spectra analysis in driver genes.** Mutations identified in the indicated genes in colorectal or liver cancers were downloaded from cBioPortal (<http://www.cbioportal.org/>). Only point mutations that resulted in a missense, nonsense or splice-site mutation were considered.

**CNV detection.** To detect copy-number variations (CNVs), BAM files were analysed for read-depth variations by CNVnator v0.2.7 (ref. 39) with a bin size of 1 kb and Control-FREEC v6.7<sup>40</sup> with a bin size of 5 kb. Highly variable regions, defined as harbouring germline CNVs in at least three control samples, were excluded from the analysis. To obtain somatic CNVs, we excluded CNVs for which there was evidence in the reference sample (blood/biopsy) of the same individual. Resulting candidate CNV regions were assessed for additional structural variants on the paired-end and split-read level through DELLY v0.3.3 (ref. 41). Based on these results, we excluded five candidate CNV regions as mapping artefacts on the read-depth level and acquired base-pair accuracy of the involved breakpoints for the other events. This also revealed the tandem orientation of the duplication events and the complex structural variation in the colon sample.

Reported gene definitions (Extended Data Table 2) are based on Ensembl v75 (GCRh37)<sup>35</sup>. Common fragile sites overlapping the events were detected using existing definitions<sup>42</sup>. LINE/SINE elements within 100 bp of the breakpoints were determined with the repeat element annotation<sup>43</sup> from the UCSC genome browser<sup>34</sup> GCRh37 (retrieved 26 October 2015).

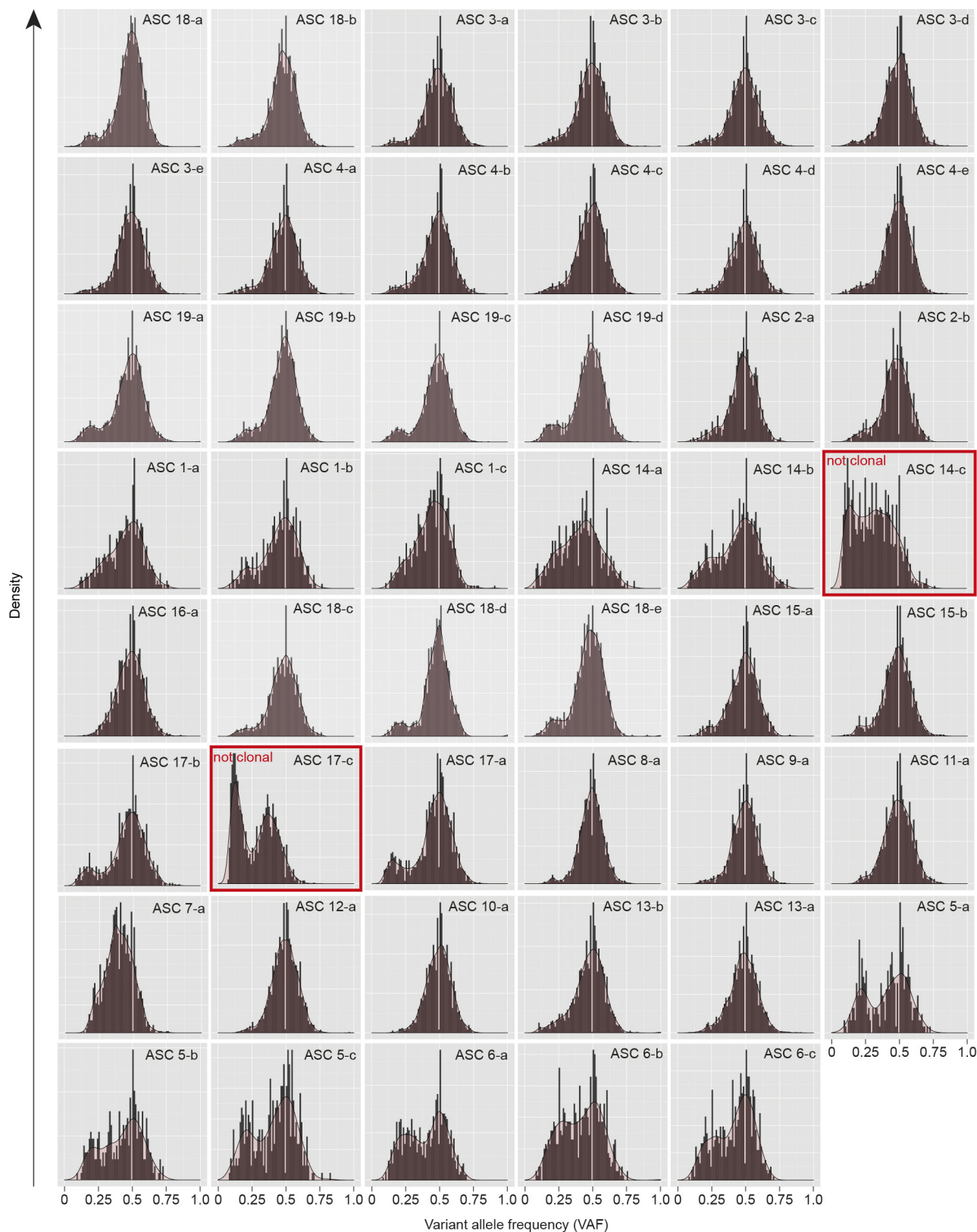
**Code availability.** All code and filtered vcf files are freely available under a MIT License at [https://wgs11.op.umcutrecht.nl/mutational\\_patterns\\_ASCs/](https://wgs11.op.umcutrecht.nl/mutational_patterns_ASCs/) and <https://github.com/CuppenResearch/MutationalPatterns/>.

27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
28. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
29. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
30. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
31. R Core Team. R: A language and environment for statistical computing; <http://www.r-project.org/> (2015).
32. Pinheiro J *et al.* nlme: Linear and Nonlinear Mixed Effects Models. <https://cran.r-project.org/web/packages/nlme/nlme.pdf> (2016).
33. ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2013).
34. Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* **43**, D670–D681 (2015).
35. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43**, D662–D669 (2015).
36. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
37. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
38. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **21**, 974–984 (2011).
39. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).
40. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
41. Le Tallec, B. *et al.* Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes. *Cell Reports* **4**, 420–428 (2013).
42. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).



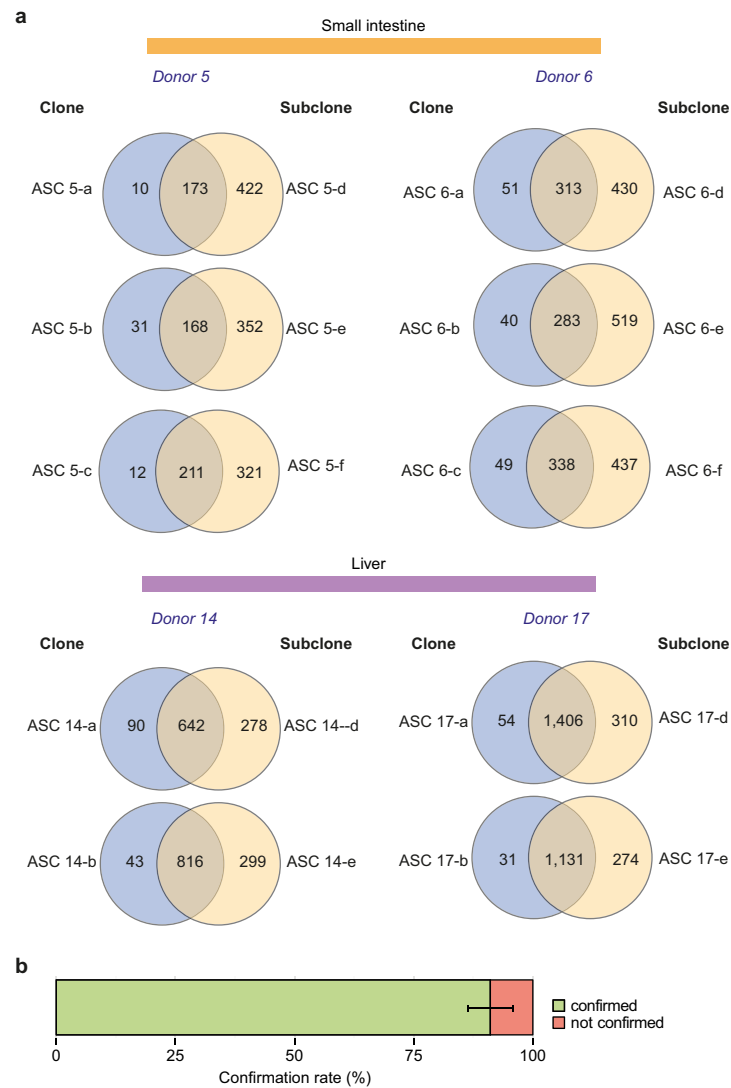
**Extended Data Figure 1 | Cataloguing somatic mutation loads in human ASCs.** **a**, Schematic overview of the experimental setup to determine somatic mutations in individual human ASCs. Colon, small intestine and liver biopsies were cultured in bulk for 1–2 week(s) before single cells were sorted and clonally expanded until enough DNA could be isolated for WGS analysis. WGS of the clonal organoid culture allows for cataloguing of somatic variants in the original ASCs that gave rise to the clonal cultures that were acquired during life and the first 7–14 days of culturing. Biopsy or blood was sequenced as a reference sample. **b**, Filter steps to obtain somatic mutations in ASCs. **c**, Number of point mutations that pass each corresponding filter step in **a** for each ASC culture of donors 5 and 6. **d**, Independent validations of mutations that were filtered out by amplicon-based resequencing. The asterisk indicates

a position that is not located in the surveyed areas of the assessed ASCs in the original experiment, which is corrected for in all analyses. **e**, Independent validations of mutations that passed all filters by amplicon-based re-sequencing. Confirmed positions are defined as those with a call in the indicated ASC with a VAF  $\geq 0.3$  and without a call in the corresponding reference sample. **f**, Qualification of unconfirmed positions based on manual inspection. True-positive positions are positions that were correctly called, but for which the VAF threshold was not met in the validation experiment. False-positive positions are positions without evidence in the validation experiment or are noisy. ‘Missed in germline’ are positions that were called in the reference sample in the validation experiment.



**Extended Data Figure 2 | Variant allele frequency distribution plot for each assessed ASC.** A distribution plot of the VAFs of all somatic mutations that remain before filtering for the VAF in filter step 6 (Extended Data Fig. 1b). Clonal heterozygous somatic mutations form a peak around VAF = 0.5. A threshold of VAF  $\geq 0.3$  was used to obtain somatic mutations that were clonal in the organoid cultures and therefore

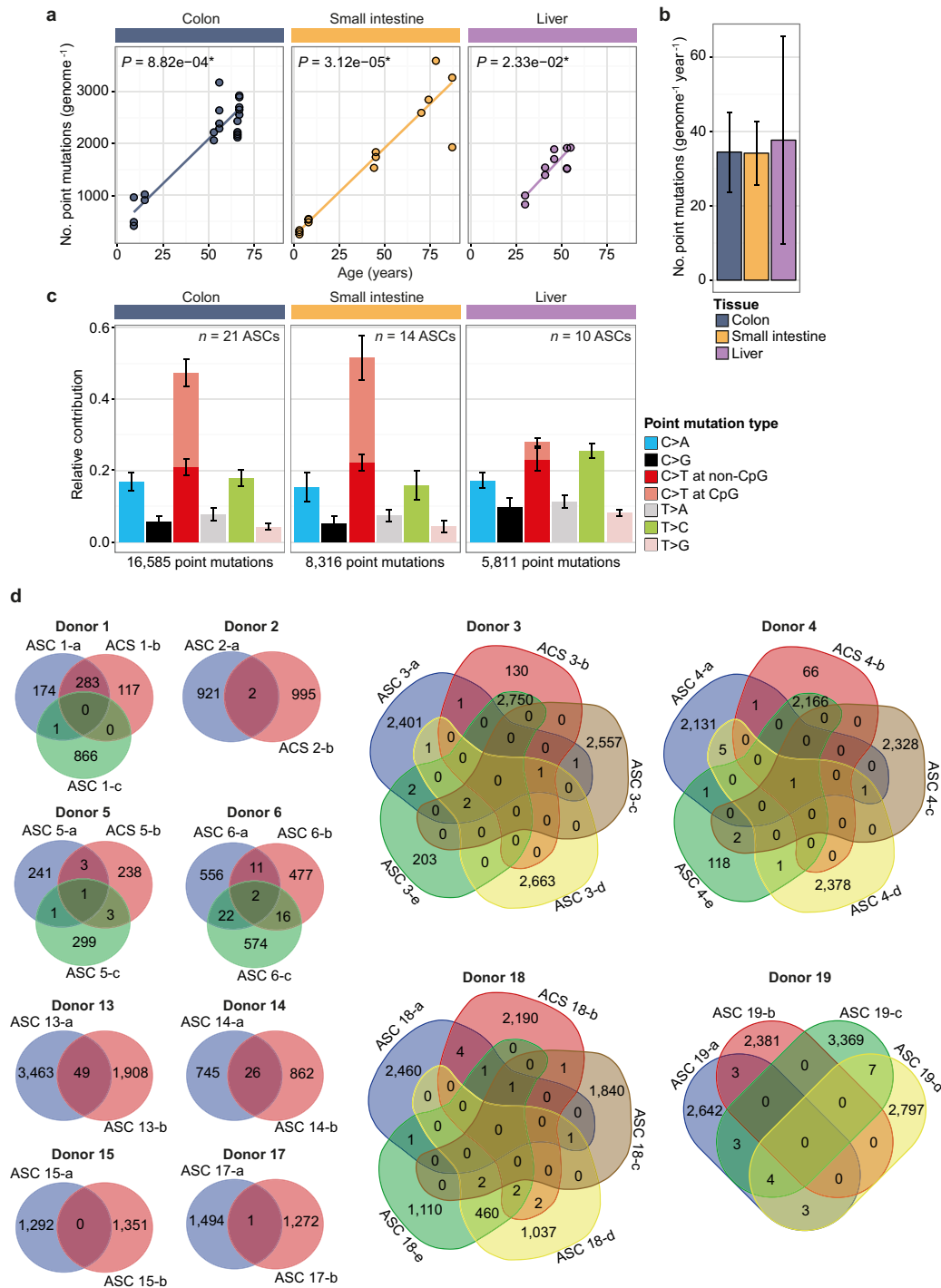
present in the original cloned ASCs (see Methods). Mutations acquired after the single ASC expansion step are subclonal (that is, not present in all cells of the clonal culture) and have lower VAFs. Two samples (donor 14, ASC 14-b and donor 17, ASC 17-c) showed a shift in the main VAF peak to the left, indicating that these cultures did not arise from a single ASC and were therefore excluded from the study.



**Extended Data Figure 3 | Confirmation rate of somatic point mutations.** **a**, Overlap of somatic point mutations between the clonal organoid cultures and corresponding subcloned cultures depicted in Extended Data Fig. 6. **b**, Confirmation rate of point mutations, which were

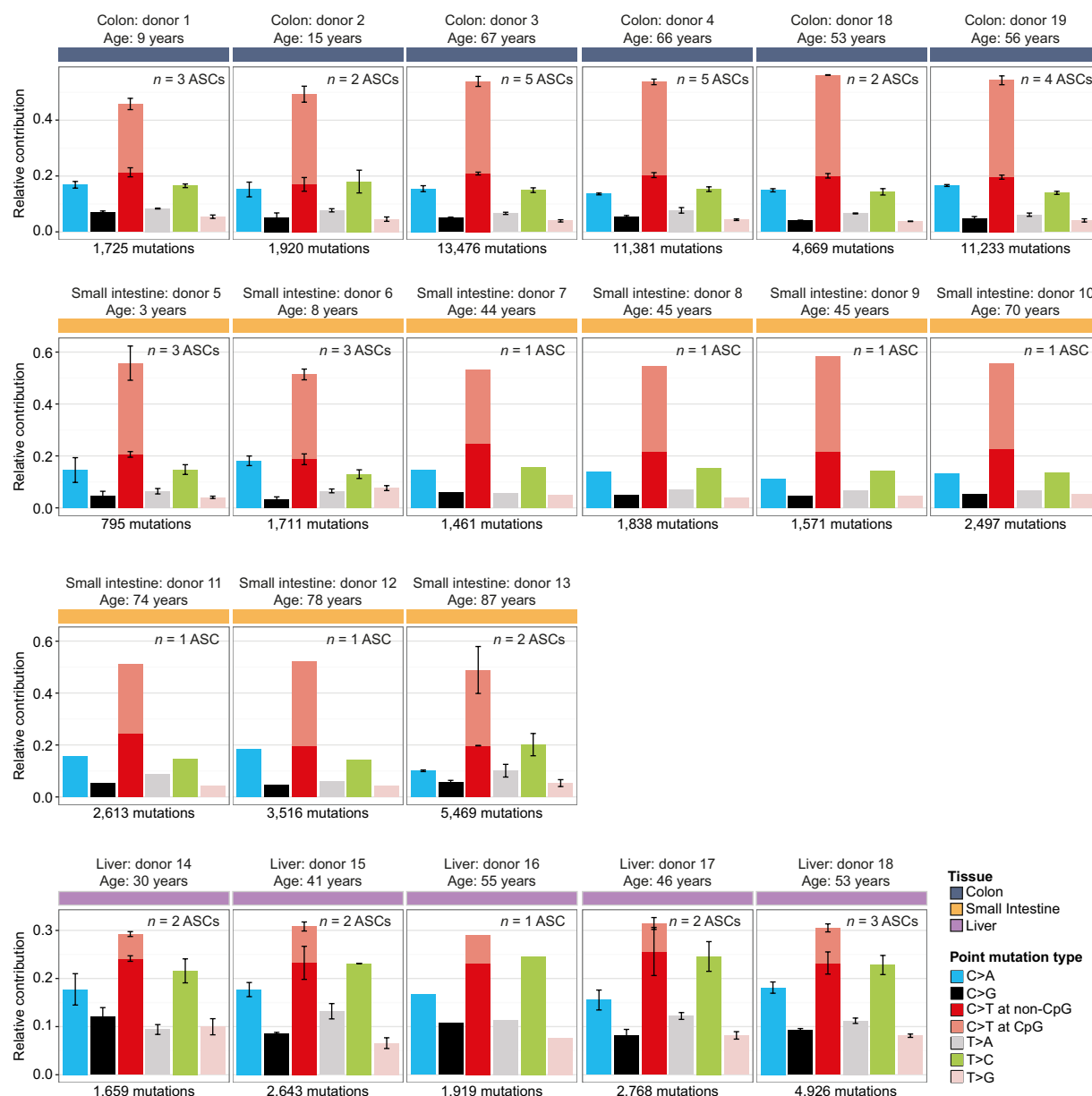
observed in the original cloned culture, in the corresponding subcloned culture. Data are represented as the mean percentage of confirmed point mutations over all clone–subclone pairs indicated in **a** ( $n = 10$ ) and error bars represent s.d.





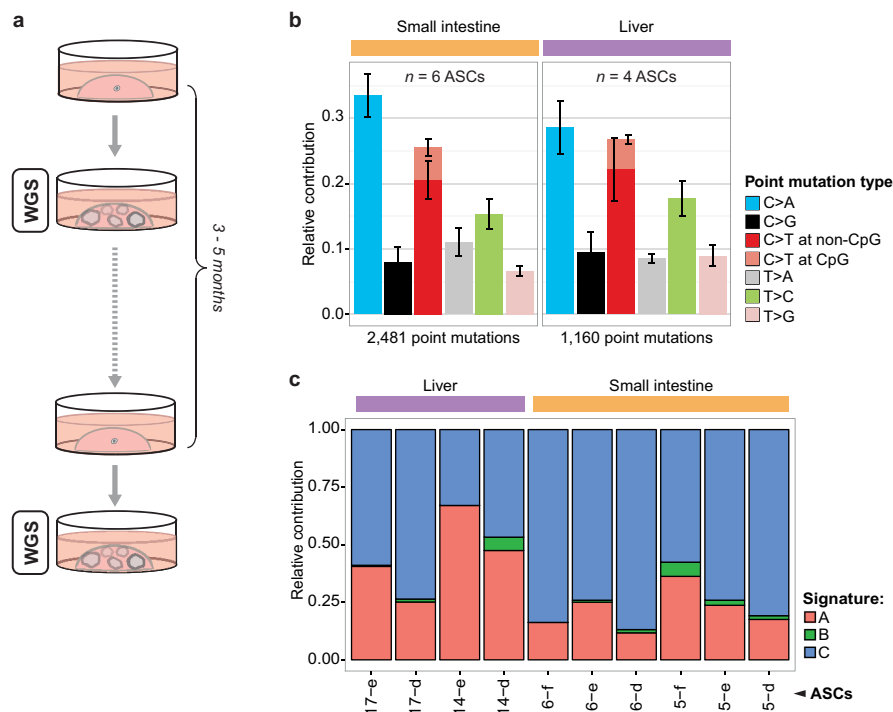
**Extended Data Figure 4 | Somatic mutation loads in consensus-surveyed area and overlap of point mutations between ASCs from the same donor.** **a**, Correlation of the number of somatic point mutations per ASC, which were observed in the genomic regions that were surveyed (for example, a base coverage of at least  $20\times$  in both the clonal culture and the reference sample; Methods) in all the ASCs, with the age of the donors per tissue indicated. This consensus-surveyed area comprises 38.2% of the non-N autosomal genome. Each data point represents a single ASC. Indicated are the  $P$  values of the age effects in the linear mixed model (two-tailed  $t$ -test) for each tissue. The sample sizes for colon, small intestine and liver are 6, 9 and 5 donors and 21, 14 and 10 ASCs, respectively. **b**, Somatic mutation accumulation rate per tissue as estimated

by the linear mixed models in **a**. Error bars represent the 95% confidence intervals of the slope estimates. **c**, Relative contribution of the indicated mutation types to the point mutation spectra in the consensus-surveyed area per tissue type. Data are represented as the mean relative contribution of each mutation type over all ASCs per tissue type ( $n = 21, 14$  and  $10$  for colon, small intestine and liver, respectively); error bars represent s.d. The total number of identified somatic point mutations per tissue is shown. **d**, Overlap of the somatic point mutations between ASCs of the same donor. The number of point mutations, observed in the total surveyed area per ASC, that are shared between the assessed ASCs of the same donor is indicated.



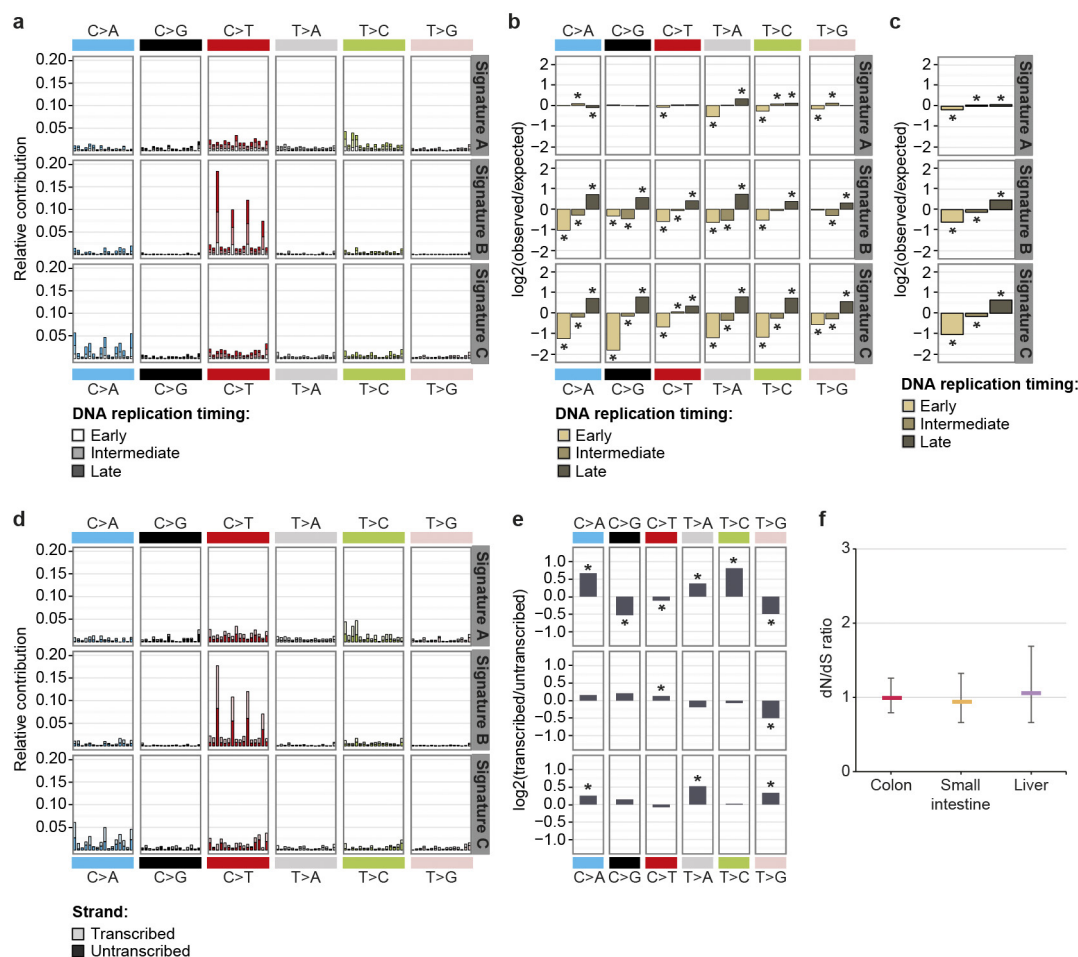
**Extended Data Figure 5 | Point-mutation spectrum per donor.** Relative contribution of the different types of point mutation to the spectrum of each donor. Data are represented as the mean relative contribution of each mutation type when multiple ASCs were measured per donor (the number  $n$

of ASC per donor is depicted for each donor) and error bars represent standard deviation. Indicated are the age of the donors, the total number of point mutations used to determine each spectrum and the tissue type.



**Extended Data Figure 6 | Mutation patterns associated with long-term *in vitro* expansion of ASCs.** **a**, Schematic overview of the experimental setup to catalogue mutations associated with the organoid culture system. Clonal small intestinal and liver organoid cultures (Extended Data Fig. 1a) were cultured for 3–5 months. A second clonal expansion step was subsequently performed, followed by WGS analysis, to catalogue all the mutations that were present in the subcloned ASCs. To obtain mutations that were specifically acquired during culturing, mutations in the original clonal cultures were subtracted from those observed in the corresponding second subcloned cultures. **b**, Relative contribution of the indicated

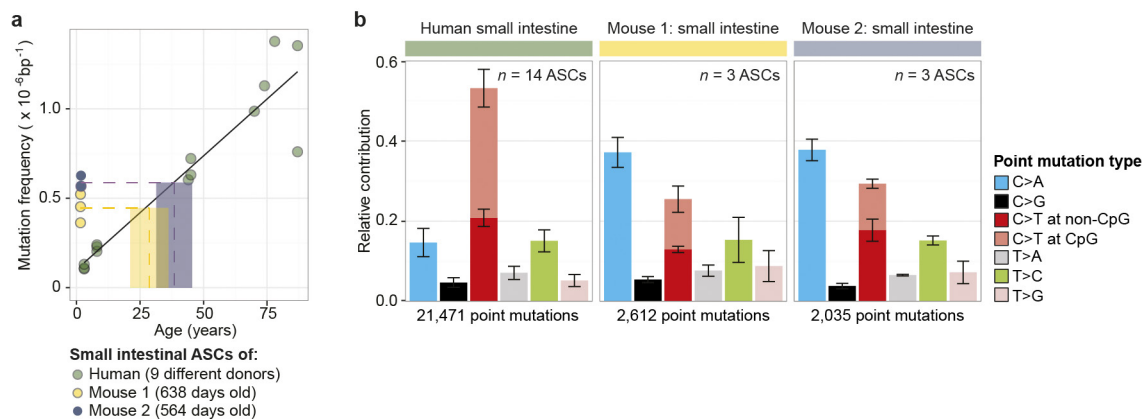
mutation types to the point mutation spectra specifically observed *in vitro* per tissue type. Data are represented as the mean relative contribution of each mutation type over all subcloned ASCs per tissue type ( $n = 6$  and 4 for small intestine and liver, respectively) and error bars represent s.d. Indicated are the total number of identified somatic point mutations, which were specifically acquired between the two clonal expansion steps indicated in **a**, per tissue. **c**, Relative contribution of the mutational signatures depicted in Fig. 2a, which explain the mutation spectra depicted in **b**.



**Extended Data Figure 7 | Non-random distribution of mutational signatures throughout the genome.** **a**, Context- and replication-timing-dependent mutation spectrum of the three mutational signatures depicted in Fig. 2a. Indicated is the contribution of each trinucleotide to the signatures (order is similar as in ref. 11), subdivided into the fraction of the trinucleotide-change present in early, intermediate or late replicating genomic regions. **b**, log<sub>2</sub> ratio of the observed and expected number of mutations per indicated base substitution (summed over all trinucleotides) in early-, intermediate- and late-replicating genomic regions for each of the signatures depicted in **a**. log<sub>2</sub> ratio indicates the effect size of the bias and asterisks indicate significant DNA-replication-timing bias ( $P < 0.05$ , binomial test). **c**, log<sub>2</sub> ratio of the total number of observed and expected mutations in early-, intermediate- and late-replicating genomic regions for each signature depicted in **a**. log<sub>2</sub> ratio indicates the effect-size of the bias

and asterisks indicate significant DNA replication timing bias ( $P < 0.05$ , binomial test). **d**, Context- and transcriptional-strand-dependent mutation spectrum of the three mutational signatures depicted in Fig. 2a. Indicated is the contribution of each trinucleotide to the signatures (order is similar to that in ref. 11), subdivided into the fraction of the trinucleotide-change present on the transcribed and untranscribed strand. **e**, log<sub>2</sub> ratio of the number of mutations on the transcribed and untranscribed strand per indicated base substitution for each signature depicted in **d**. log<sub>2</sub> ratio indicates the effect size of the bias and asterisks indicate significant transcriptional strand bias ( $P < 0.05$ , binomial test). **f**, The dN/dS ratio for all protein-coding somatic point mutations observed in all ASCs per tissue type. Error bars indicate 95% confidence intervals (likelihood ratio test).





**Extended Data Figure 8 | Comparison of mutation loads between intestinal ASCs derived from human and mouse.** **a**, Mutation frequency in mouse intestinal ASCs is compared to the linear fit, describing the relationship between the mutation frequency in human intestinal ASCs and age of the donor. Indicated by the dotted lines are the mean mutation frequencies over all ASCs per mouse ( $n = 3$ ) and the corresponding age of

human linear fit. **b**, Relative contribution of the indicated mutation types to the point mutation spectra for all assessed human intestinal ASCs and for each mouse. Data are represented as the mean relative contribution of each mutation type over all the ASCs per indicated category ( $n = 14$ , 3 and 3 for human, mouse 1 and mouse 2, respectively), error bars indicate s.d.

**Extended Data Table 1 | Overview of somatic point mutations detected in ASCs**

ASC	Donor	Age (years)	Gender	Tissue	Surveyed genome (%) <sup>*</sup>	No. point mutations <sup>†</sup>
1-a	1	9	Female	Colon	93.8	458
1-b	1	9	Female	Colon	91.2	400
1-c	1	9	Female	Colon	97.6	867
2-a	2	15	Male	Colon	96.8	923
2-b	2	15	Male	Colon	96.8	997
18-a	18	53	Male	Colon	98.5	2,468
18-b	18	53	Male	Colon	98.1	2,201
19-a	19	56	Male	Colon	97.7	2,655
19-b	19	56	Male	Colon	97.1	2,384
19-c	19	56	Male	Colon	98.2	3,383
19-d	19	56	Male	Colon	97.8	2,811
4-a	4	66	Female	Colon	90.7	2,140
4-b	4	66	Female	Colon	95.3	2,234
4-c	4	66	Female	Colon	95.7	2,332
4-d	4	66	Female	Colon	93.9	2,386
4-e	4	66	Female	Colon	96.1	2,289
3-a	3	67	Female	Colon	91.8	2,409
3-b	3	67	Female	Colon	91.8	2,882
3-c	3	67	Female	Colon	91.9	2,561
3-d	3	67	Female	Colon	92.0	2,667
3-e	3	67	Female	Colon	92.0	2,957
5-a	5	3	Female	Small intestine	89.0	246
5-b	5	3	Female	Small intestine	85.5	245
5-c	5	3	Female	Small intestine	88.5	304
6-a	6	8	Female	Small intestine	97.1	591
6-b	6	8	Female	Small intestine	93.5	506
6-c	6	8	Female	Small intestine	96.2	614
7-a	7	44	Male	Small intestine	91.1	1,461
8-a	8	45	Male	Small intestine	95.5	1,838
9-a	9	45	Male	Small intestine	93.9	1,571
10-a	10	70	Female	Small intestine	94.8	2,497
11-a	11	74	Male	Small intestine	87.3	2,613
12-a	12	78	Female	Small intestine	95.6	3,516
13-a	13	87	Male	Small intestine	97.7	3,512
13-b	13	87	Male	Small intestine	97.0	1,957
14-a	14	30	Male	Liver	81.3	771
14-b	14	30	Male	Liver	85.2	888
15-a	15	41	Female	Liver	93.5	1,292
15-b	15	41	Female	Liver	95.1	1,351
17-a	17	46	Female	Liver	79.4	1,495
17-b	17	46	Female	Liver	73.7	1,273
18-c	18	53	Male	Liver	97.9	1,845
18-d	18	53	Male	Liver	98.5	1,504
18-e	18	53	Male	Liver	98.2	1,577
16-a	16	55	Male	Liver	97.5	1,919

<sup>\*</sup>Percentage of the non-N autosomal genome with  $\geq 20\times$  coverage in both ASC culture and reference sample.

<sup>†</sup>Number of somatic point mutations detected within surveyed genome.

**Extended Data Table 2 | Identified somatic structural variations in ASCs***Copy Number Variants*

Sample	Tissue	Chr	Start	Stop	Size	Type	No. genes	Fragile site	Microhomology	Genes at breakpoint	LINE/SINE
ASC 14-a	Liver	3	94,491,729	95,651,811	1,160,082	gain	5	-	5 bp	-	L1MC1
ASC 14-a	Liver	3	111,726,406	113,471,637	1,745,231	gain	46	-	2 bp	TAGLN3 ATP6V1A	L1MC1 L1M5
ASC 16-a	Liver	9	50,763,759	141,213,431	90,449,672	gain	1,472	-	NA	NA	NA
ASC 18-e	Liver	7	132,751,706	133,009,202	257,496	gain	2	-	0 bp	CHCHD3 EXOC4	MIR L1PA6
ASC 18-d	Liver	5	59,125,105	59,718,364	593,259	loss	1	-	0 bp	PDE4D PDE4D	- L1PA6
ASC 8-a	Small intestine	5	3,815,936	3,908,819	92,883	loss	0	-	2 bp	-	-
ASC 11-a	Small intestine	2	205,420,067	205,511,877	91,810	loss	1	FRA2I	1 bp	PARD3B PARD3B	AluSx L1ME3B
ASC 13-a	Small intestine	11	63,974,352	66,222,668	2,248,316	loss	163	-	3 bp	FERMT3	- L1M4b
ASC 13-b	Small intestine	1	5,878,566	6,321,750	443,184	loss	13	FRA1A	1 bp	-	THE1B
ASC 1-c	Colon	3	60,700,662	61,199,328	498,666	loss	4	FRA3B	1 bp	FHIT FHIT	L1PA3 L1PA3
ASC 3-c	Colon	13	0	115,169,878	115,169,878	gain	1,217	-	NA	NA	NA
ASC 4-b&e	Colon	14	102,805,595	104,172,376	1,366,781	loss	57	-	NA	NA	NA
ASC 4-b&e	Colon	17	2,429,169	2,572,747	143,578	loss	5	-	CTTG ins	- PAFAH1B1	AluJo AluSq
ASC 4-b&e	Colon	17	2,634,433	2,927,007	292,574	loss	4	-	NA	NA	NA

*Unbalanced Translocations*

Sample	Tissue	Chr (1)	Position (1)	Chr (2)	Position (2)	Type	No. genes	Fragile site	Microhomology	Genes at breakpoint	LINE/SINE
ASC 4-b&e	Colon	14	102,805,595	17	2,634,145	translocation	NA	-	4 bp	ZNF839	-
ASC 4-b&e	Colon	14	104,172,376	18	18,518,987	translocation	NA	-	0 bp	XRCC3	- ALR Alpha
ASC 4-b&e	Colon	17	2,927,007	18	18,518,987	translocation	NA	-	0 bp	RAF1GAP2	L1PA8 ALR Alpha

No. genes, number of genes overlapping the event; fragile site, common fragile sites overlapping the event; microhomology, number of bases of microhomology observed at breakpoints; genes at breakpoint, gene bodies affected by the breakpoint; LINE/SINE elements, observed elements within 100 bp of the breakpoint.

# Formation of new chromatin domains determines pathogenicity of genomic duplications

Martin Franke<sup>1,2\*</sup>, Daniel M. Ibrahim<sup>1,2,3\*</sup>, Guillaume Andrey<sup>1</sup>, Wibke Schwarzer<sup>4</sup>, Verena Heinrich<sup>2,5</sup>, Robert Schöpflin<sup>5</sup>, Katerina Kraft<sup>1,2</sup>, Rieke Kempfer<sup>1</sup>, Ivana Jerković<sup>1,2</sup>, Wing-Lee Chan<sup>2</sup>, Malte Spielmann<sup>1,2</sup>, Bernd Timmermann<sup>6</sup>, Lars Wittler<sup>7</sup>, Ingo Kurth<sup>8,9</sup>, Paola Cambiaso<sup>10</sup>, Orsetta Zuffardi<sup>11</sup>, Gunnar Houge<sup>12</sup>, Lindsay Lambie<sup>13</sup>, Francesco Brancati<sup>14,15</sup>, Ana Pombo<sup>3,16</sup>, Martin Vingron<sup>5</sup>, Francois Spitz<sup>4</sup> & Stefan Mundlos<sup>1,2,3,17</sup>

Chromosome conformation capture methods have identified subchromosomal structures of higher-order chromatin interactions called topologically associated domains (TADs) that are separated from each other by boundary regions<sup>1,2</sup>. By subdividing the genome into discrete regulatory units, TADs restrict the contacts that enhancers establish with their target genes<sup>3–5</sup>. However, the mechanisms that underlie partitioning of the genome into TADs remain poorly understood. Here we show by chromosome conformation capture (capture Hi-C and 4C-seq methods) that genomic duplications in patient cells and genetically modified mice can result in the formation of new chromatin domains (neo-TADs) and that this process determines their molecular pathology. Duplications of non-coding DNA within the mouse *Sox9* TAD (intra-TAD) that cause female to male sex reversal in humans<sup>6</sup>, showed increased contact of the duplicated regions within the TAD, but no change in the overall TAD structure. In contrast, overlapping duplications that extended over the next boundary into the neighbouring TAD (inter-TAD), resulted in the formation of a new chromatin domain (neo-TAD) that was isolated from the rest of the genome. As a consequence of this insulation, inter-TAD duplications had no phenotypic effect. However, incorporation of the next flanking gene, *Kcnj2*, in the neo-TAD resulted in ectopic contacts of *Kcnj2* with the duplicated part of the *Sox9* regulatory region, consecutive misexpression of *Kcnj2*, and a limb malformation phenotype. Our findings provide evidence that TADs are genomic regulatory units with a high degree of internal stability that can be sculptured by structural genomic variations. This process is important for the interpretation of copy number variations, as these variations are routinely detected in diagnostic tests for genetic disease and cancer. This finding also has relevance in an evolutionary setting because copy-number differences are thought to have a crucial role in the evolution of genome complexity.

*SOX9* is a developmental transcription factor with important functions in chondrocyte differentiation and male sex determination<sup>7</sup>. The *SOX9* locus is a genomic region that has been linked to various human diseases with a broad range of phenotypes<sup>6,8,9</sup>. We investigated how genomic duplications in this region affect higher-order chromatin organization and, in particular, the formation of TADs. Hi-C, a technology to quantify chromatin contacts genome-wide, shows a compartmentalization of the human locus in two major TADs, one containing *SOX9* (referred to hereafter as *SOX9* TAD), the other

containing the two potassium channels *KCNJ2* and *KCNJ16* (referred to hereafter as *KCNJ* TAD) (Fig. 1a)<sup>2</sup>. The large gene desert corresponding to the *SOX9* TAD has been shown to contain multiple regulatory elements and human-disease-related sites (Fig. 1a)<sup>9</sup>. Duplications including a region 0.5 megabases (Mb) upstream of *SOX9* (referred to as the RevSex region) lead to female-to-male sex reversal (Fig. 1a), indicating that the RevSex region contains a critical regulator of *SOX9* expression in the developing gonads<sup>6,10</sup>. Surprisingly, duplications that include the RevSex region but extend further upstream towards the neighbouring genes *KCNJ2* and *KCNJ16* have no effect on sexual development. Instead, they result in Cooks syndrome, a congenital limb malformation characterized by aplasia of nails and short digits<sup>8</sup>. We identified a family (mother and daughter) with a third type of duplication that includes the RevSex region and the entire gene desert upstream of *SOX9*, but not the *KCNJ2* and *KCNJ16* genes (Fig. 1a). In spite of complete overlap with the reported sex reversal duplications and large parts of the Cooks syndrome duplication, carriers of this variant are phenotypically normal. According to the Hi-C profile, all sex reversal associated duplications are located within the *SOX9* TAD, whereas the duplication with no phenotype and also the Cooks syndrome duplications extend into the neighbouring *KCNJ* TAD thereby spanning two TADs and their boundary. We refer to these two types of duplications as intra-TAD and inter-TAD, respectively.

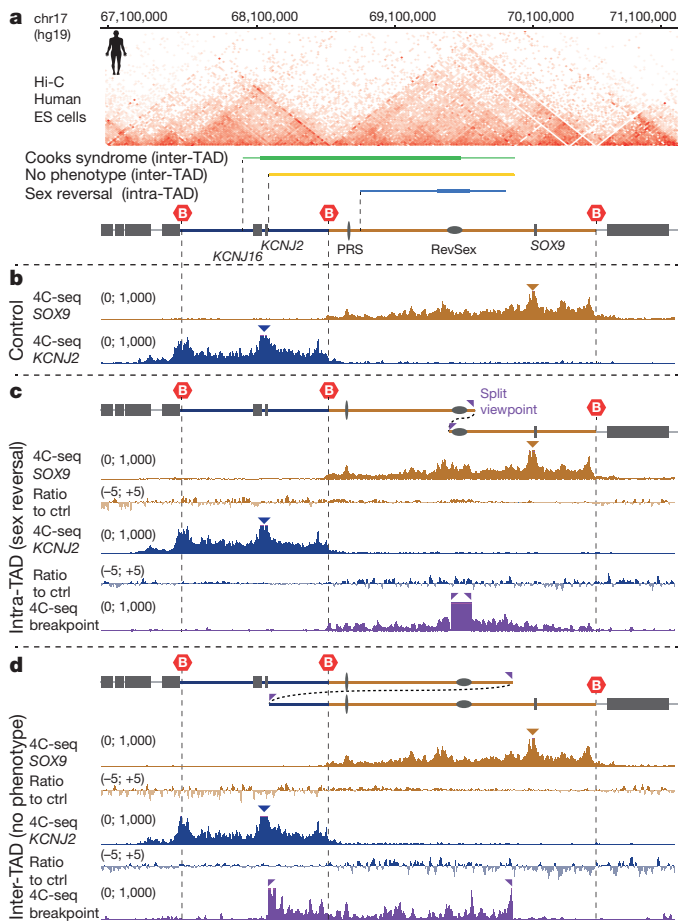
To investigate the difference between the intra-TAD (sex reversal) and the inter-TAD (no phenotype) duplications, we performed 4C-seq of patient fibroblasts and compared them to controls. In control cells, the *SOX9* and the *KCNJ2* viewpoint showed an interaction profile restricted to their Hi-C predicted TADs (Fig. 1b). 4C-seq from a patient with sex reversal and a 150 kilobase (kb) intra-TAD duplication showed a slight increase with the duplicated part from *SOX9*, but an overall unchanged profile from both viewpoints (Fig. 1c). Similarly, no change in the configuration of the interaction profiles was observed with the larger inter-TAD (no phenotype) duplication (Fig. 1d). To identify the region specifically contacted by the duplicated part, we used the unique sequence created by the duplication breakpoints to perform allele-specific 4C-seq. In the intra-TAD duplication, we observed high interaction with the entire *SOX9* TAD (purple track, Fig. 1c). Similar results were obtained with fibroblasts from a second individual with a 470 kb intra-TAD duplication (Extended Data Fig. 1). By contrast, the interaction profile generated from the inter-TAD duplication breakpoint was restricted to the duplicated region, without contacting

<sup>1</sup>Max Planck Institute for Molecular Genetics, RG Development & Disease, 14195 Berlin, Germany. <sup>2</sup>Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, 13353 Berlin, Germany. <sup>3</sup>Berlin Institute of Health, 10117 Berlin, Germany. <sup>4</sup>Developmental Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. <sup>5</sup>Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, 14195 Berlin, Germany. <sup>6</sup>Max Planck Institute for Molecular Genetics, Sequencing Core Facility, 14195 Berlin, Germany.

<sup>7</sup>Max Planck Institute for Molecular Genetics, Department Developmental Genetics, 14195 Berlin, Germany. <sup>8</sup>Institute of Human Genetics, Jena University Hospital, 07743 Jena, Germany. <sup>9</sup>Institute of Human Genetics, Uniklinik RWTH Aachen, 52074 Aachen, Germany. <sup>10</sup>Bambino Gesù Children's Hospital-IRCCS, 00165 Rome, Italy. <sup>11</sup>Department of Molecular Medicine, University of Pavia, 27100 Pavia, Italy. <sup>12</sup>Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, 5021 Bergen, Norway. <sup>13</sup>Division of Human Genetics, National Health Laboratory Service, University of the Witwatersrand, 2000 Johannesburg, South Africa. <sup>14</sup>Department of Life, Health and Environmental Sciences, University of L'Aquila, 67100 L'Aquila, Italy. <sup>15</sup>Istituto Dermatologico dell'Immacolata (IDI) IRCCS, 00167 Rome, Italy. <sup>16</sup>Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, 13125 Berlin-Buch, Germany.

<sup>17</sup>Berlin-Brandenburg Center for Regenerative Therapies (BCRT), Charité Universitätsmedizin Berlin, 13353 Berlin, Germany. \*These authors contributed equally to this work.





**Figure 1 | 4C-seq from human fibroblasts with intra-TAD and inter-TAD duplications.** **a**, SOX9 locus with Hi-C interaction profile (H1 embryonic stem cells<sup>2</sup>). Schematic shows KCNJ TAD (blue), SOX9 TAD (brown) and TAD boundaries (red hexagons). Ovals indicate regions associated with sex reversal (RevSex) or Pierre Robin syndrome (PRS), human duplications with corresponding phenotypes are indicated above (line thickness indicates maximum/minimum size). **b**, 4C-seq from control skin fibroblasts using SOX9 and KCNJ2 as viewpoints (triangles). **c**, 4C-seq from fibroblasts of an individual with sex reversal and intra-TAD duplication shows no change in overall interaction profile (below, log<sub>2</sub> ratio to control). 4C-seq from duplication breakpoint (below, in purple) shows increased interaction with the duplicated part and the rest of the SOX9 TAD. **d**, 4C-seq from fibroblasts of individual with no phenotype and inter-TAD duplication. Note unchanged interaction profiles from SOX9 and KCNJ2 viewpoints. 4C-seq from duplication breakpoint shows interactions restricted to the duplicated region. All reads mapped to a wild-type genome (resulting in split viewpoint for duplication breakpoints). See Methods for sample collection.

the SOX9, KCNJ2 or KCNJ16 genes (purple track, Fig. 1d). To further investigate allele-specific interactions with the different viewpoints, we performed whole-genome sequencing of the patient sample with the inter-TAD duplication and determined if certain single nucleotide polymorphisms (SNPs) were contacted by one viewpoint and not the other. We identified several SNPs that were only contacted by the duplication specific viewpoint but not by the SOX9 or KCNJ2 viewpoints, suggesting that the duplicated region contacted primarily itself, thereby forming a separate domain (Extended Data Fig. 2 and Methods).

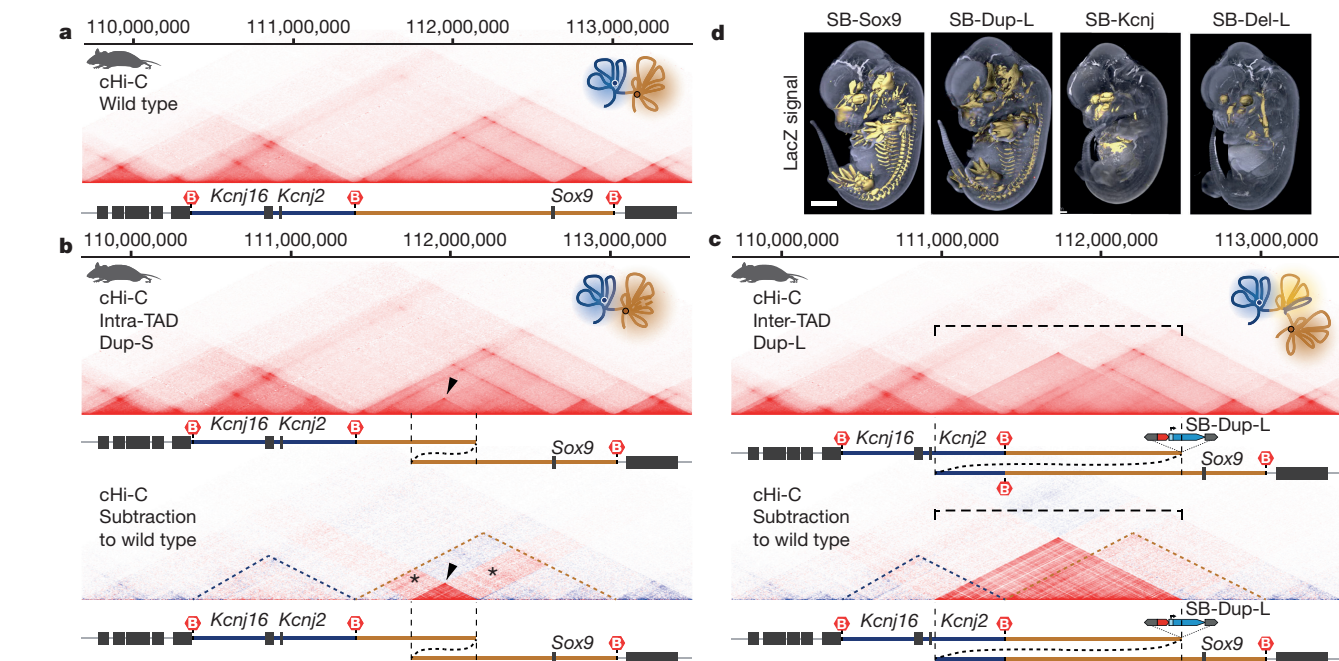
To investigate the effect of these duplications on TAD structure in more detail, we re-engineered the human duplications in mice and performed capture Hi-C (cHi-C) and 4C-seq of the extended Sox9 locus from wild-type and mutant limb buds at embryonic day 12.5 (E12.5) (Fig. 2). We used CRISPR/Cas9 genome editing to create mice with a 400 kb intra-TAD duplication (Dup-S), equivalent to a

published patient with sex reversal<sup>6</sup>. Heterozygous Dup-S mice were phenotypically normal. This is probably explained by the finding that a known mouse gonadal enhancer, TESCO, resides outside of the duplication<sup>11</sup>, and the human sex-reversal-associated regions could so far not be linked to Sox9 regulation in mouse<sup>10</sup>. The larger 1.6 Mb inter-TAD duplication (Dup-L), equivalent to the individuals with no phenotype, was generated using the Cre/loxP system and trans-allelic recombination.

In agreement with the human data, our data show that the mouse locus is subdivided into a Sox9 TAD, containing Sox9 as the only gene, and a Kcnj TAD with Kcnj2 and Kcnj16, separated by a boundary located in the gene desert (Fig. 2a). cHi-C from mice with the intra-TAD duplication (Dup-S) showed no change in the overall TAD structure (Fig. 2b). Subtraction of the wild-type profile from the Dup-S profile demonstrated increased interaction of the duplicated part with itself, but also with the rest of the Sox9 TAD including Sox9. By contrast, cHi-C from the inter-TAD duplication (Dup-L) showed a new interaction domain covering the duplicated parts of the Kcnj and Sox9 TADs (Fig. 2c). The subtraction profile showed a strong increase in interaction in comparison to wild type, but only within the duplicated region (thus excluding Sox9, Kcnj2 and Kcnj16) and no changes in the adjacent TADs. This suggested that the new interaction domain was isolated from the neighbouring Sox9 and Kcnj TADs. 4C-seq from outside of the duplication (Sox9 and Kcnj2) confirmed the isolation of this domain by showing wild-type interaction profiles and no ectopic contacts (Extended Data Fig. 3). 4C-seq from the lacZ reporter located at the duplication breakpoint corroborated our other findings, showing interaction only with the duplicated region (Extended Data Fig. 3b). In agreement with the observed isolation of the duplication domain, Dup-L mice were phenotypically normal and fertile.

Thus, the two types of duplications differ fundamentally in their effect on chromatin organization in spite of the fact that they overlap to a large degree. The intra-TAD duplication (resulting in sex reversal) has no effect on the overall TAD structure, whereas the larger inter-TAD duplication (resulting in no phenotype in humans and mice) resulted in the generation of a new chromatin domain that we refer to as a 'neo-TAD'. The insulation of the neo-TAD prevents interaction with the neighbouring Sox9, Kcnj2 or Kcnj16 genes and explains the lack of pathology in humans and in mice that carry this type of duplication.

The presence of the lacZ reporter in the middle of the Dup-L duplication also allowed us to investigate the regulatory potential of the neo-TAD by analysing the lacZ expression pattern in Dup-L mutants and compare it to lacZ reporters inserted in the Kcnj and Sox9 TADs (Sleeping Beauty (SB) transposons carrying lacZ; SB-Kcnj and SB-Sox9 alleles; Fig. 2d and Extended Data Fig. 4). SB-Sox9 gave a strong signal essentially recapitulating the endogenous Sox9 expression, whereas SB-Kcnj showed a restricted signal in the developing jaw and nose. The LacZ staining obtained in Dup-L was very similar to that of SB-Sox9, suggesting that the regulatory sequence contained in the neo-TAD was functional and sufficient to recapitulate most of the Sox9 expression domains, but not that of Kcnj2. This was further supported by the analysis of a deletion (Del-L allele, Extended Data Fig. 4) in which the corresponding region had been removed by trans-allelic recombination. LacZ staining in Del-L mice showed a complete loss of the Sox9, but a preservation of the Kcnj2 expression domains (Fig. 2d). Thus, the regulatory region in the neo-TAD was able to drive expression in a tissue-specific Sox9 pattern. As predicted by reporter assays and histone modification data, the region contains a large number of regulatory sites (Extended Data Fig. 5a) that, as shown by the lacZ reporter in Del-L mice, control the major part of Sox9 expression. Interestingly, comparison of the interaction profiles of the Kcnj2 and Sox9 viewpoints in wild type with the profile from the lacZ viewpoint in Dup-L revealed a similar pattern and peak distribution of 4C-seq profiles, suggesting that the newly formed domain folded in a structured manner retaining its wild-type interaction characteristics (Extended Data Fig. 5).

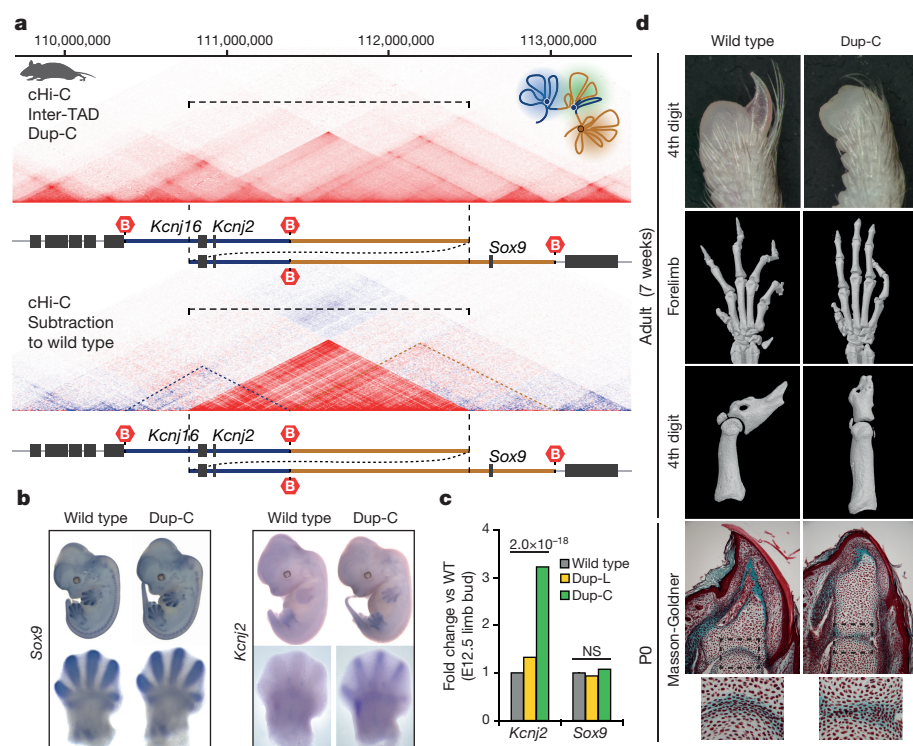


**Figure 2 | Capture Hi-C from intra-TAD and inter-TAD duplications in mice.** **a**, Capture Hi-C (cHi-C) of the *Sox9* locus from wild-type E12.5 limb buds showing separation in two TADs. Schematic on the top right depicts chromatin folding in *Kcnj* (blue) and *Sox9* (brown) TAD. **b**, cHi-C of intra-TAD duplication in Dup-S mice and subtraction map relative to wild type (below) shows an increase in interaction frequency of the duplicated region with itself (arrow) and the *Sox9* TAD (asterisks). **c**, cHi-C of inter-TAD duplication in Dup-L mice and subtraction map (below)

shows formation of a neo-TAD (brackets). *Kcnj* and *Sox9* TAD indicated by dashed lines. All coordinates chr11 (mm9). See Methods for sample collection. **d**, Optical projection tomography (OPT) of LacZ staining with insertion of *lacZ* reporter in *Sox9* TAD (SB-Sox9), at the duplication breakpoint (SB-Dup-L), in *Kcnj* TAD (SB-Kcnj), and in a deletion (SB-Del-L) corresponding to Dup-L (see Extended Data Fig. 4). Note recapitulation of *Sox9* pattern in Dup-L and loss in Del-L. Scale bar, 1 mm.

We hypothesized that an extension of the duplication towards the next flanking genes *Kcnj2* and *Kcnj16*, as reported in human Cooks syndrome<sup>8</sup>, would result in incorporation of these genes in the neo-TAD. We generated this inter-TAD duplication (Dup-C) by *trans*-allelic recombination (Fig. 3). cHi-C of E12.5 Dup-C limb buds showed, like in the Dup-L mutant, a new chromatin domain corresponding to the

duplicated region (Fig. 3a). The subtraction showed a strong increase in interaction between the duplicated parts of the *Sox9* and *Kcnj* TADs including the *Kcnj2* and *Kcnj16* genes, but not with *Sox9*. Ectopic interaction of *Kcnj2* with parts of the *Sox9* TAD was confirmed by 4C-seq from mouse limb buds and by analysing fibroblasts from a Cooks syndrome patient (Extended Data Fig. 6). To investigate the effects



**Figure 3 | Capture Hi-C, gene expression and phenotype in inter-TAD Cooks syndrome duplication.** **a**, cHi-C and subtraction map relative to wild type (below) of inter-TAD duplication in Dup-C E12.5 limb buds showing formation of new interaction domain (neo-TAD, bracket) including *Kcnj2* and *Kcnj16*. *Kcnj* and *Sox9* TAD indicated by dashed lines. Coordinates chr11 (mm9). **b**, Whole mount *in situ* hybridization for *Sox9* and *Kcnj2* at E12.5. Note *Sox9*-like expression of *Kcnj2* in Dup-C digits,  $n = 5$ . **c**, RNA-seq from E12.5 limb buds shows upregulation of *Kcnj2* in Dup-C mice only (Benjamini–Hochberg adjusted  $P$  value,  $n = 2$ , see Methods). **d**, Phenotypic comparison of wild type and the Dup-C mutant. Note (top to bottom) absence of claw, abnormal shape and positioning of phalanges (micro-CT), hypoplasia of nail bed and abnormal distal phalanx (Masson–Goldner). Enlargement (dashed boxes) showing joint fusion in Dup-C mutants. Phenotype is fully penetrant ( $n = 30$ ).

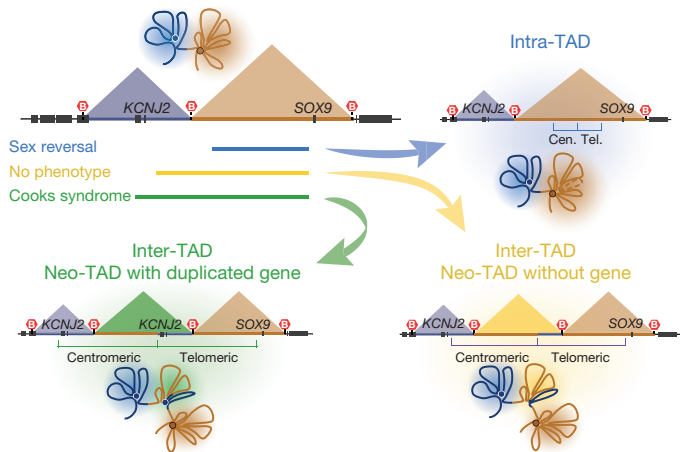


of these ectopic contacts on gene expression, we performed whole-mount *in situ* hybridization and found *Kcnj2* to be expressed in the digit anlagen, with a pattern similar to *Sox9* (Fig. 3b). RNA sequencing (RNA-seq) expression analysis of Dup-C limb buds at E12.5 and E17.5 confirmed the upregulation of *Kcnj2*, whereas other genes around the locus stayed unchanged, in particular *Sox9*, but also *Kcnj16* (Fig. 3c and Extended Data Fig. 7). Thus, the inclusion of *Kcnj2* in the neo-TAD resulted in its activation by regulatory elements that originally belonged to the *Sox9* TAD. In contrast to *Kcnj2*, *Kcnj16* was not responsive to ectopic activation. 4C-seq profiles in human fibroblasts with *KCNJ2* and *KCNJ16* as viewpoints showed slight differences in local intensity but generally the same size of ectopic interaction (Extended Data Fig. 6), indicating that a certain permissiveness or specificity must be present for promoter activation in this setting.

These new contacts and the associated misexpression of *Kcnj2* were accompanied by major phenotypic changes. Heterozygous Dup-C mice showed a limb malformation phenotype at birth closely resembling Cooks syndrome<sup>8</sup>, highlighted by the absence or severe hypoplasia of all claws or nails (Fig. 3d). Micro-computed tomography ( $\mu$ CT) and histology demonstrated an abnormal shape and size of the distal phalanges, which were fixed in a straight position due to a malformed terminal phalanx and a partly fused interphalangeal joint. To rule out that the Cooks phenotype was merely produced by increased gene dosage of *Kcnj2*, we created intra-TAD duplications that included *Kcnj2* and *Kcnj16* (Dup-K<sub>1</sub> and Dup-K<sub>2</sub>) (Extended Data Fig. 4). These mice were normal and had no digit phenotype. In addition, we created a second Cooks allele (Dup-C<sub>2</sub>) that included *Kcnj2* in the duplication, but not *Kcnj16*. These mice showed the typical Cooks syndrome phenotype. Taken together, our data suggest misexpression of *Kcnj2* as the cause for Cooks syndrome. The *Drosophila melanogaster* *KCNJ2* homologue *Irk2*, an inwardly rectifying K<sup>+</sup> channel, has been shown to have additional functions in development via the *dpp* (bone morphogenetic protein, BMP) pathway<sup>12</sup>. Mutations in components of the BMP pathway are a major cause of abnormalities in digits and joints<sup>13</sup>, providing a possible connection to the observed pathology in Cooks syndrome.

The Dup-C duplication resulted in the inclusion of *Kcnj2* in the neo-TAD and its positioning next to *Sox9* regulatory domain without an intervening boundary. We hypothesized that a similar effect might be achieved by removing the boundary between the *Sox9* and *Kcnj* TADs. To test this hypothesis, we deleted using CRISPR/Cas9 the predicted boundary region (Bor), a small (18 kb) region containing conserved CTCF binding sites. Mice with this deletion ( $\Delta$ Bor) had no apparent phenotype. cHi-C analyses of homozygous  $\Delta$ Bor limb buds showed an increase of interaction but no fusion of the *Kcnj* and *Sox9* TADs (Extended Data Fig. 8a). This ectopic interaction resulted in the upregulation, but no site-specific misexpression of *Kcnj2* (Extended Data Fig. 8b). Thus, deletion of the boundary resulted in ectopic contacts, as previously reported by others *in vitro*<sup>1,14</sup>, but the overall TAD structure remained unchanged. Similar results were obtained when deleting the boundaries in the Dup-L duplication (Extended Data Fig. 8). We observed increased interaction between the TADs, but overall the neo-TAD remained stable, indicating that the neo-TAD behaved like a 'regular' TAD. The importance of boundaries in restricting chromatin interactions was highlighted previously by deletions at the *Epha4* locus<sup>3</sup>. However, in these experiments large portions of the adjacent TADs were deleted together with the boundaries, thereby disrupting the overall TAD structure. Our present data indicate that other factors, such as additional CTCF sites and loops within TADs<sup>14,15</sup> contribute to TAD stability. The deletion of a boundary alone has therefore no major consequences, whereas larger deletions result in a re-organization of the locus enabling new contacts.

Duplications are generally thought to confer their phenotypic effect through an increase in gene dosage, but often the observed phenotype cannot be explained by alterations in gene dosage. Our data show how duplications can have different effects on higher-order chromatin structure, depending on their size and position. Duplications that are



**Figure 4 | Duplication-induced effects on chromatin organization and phenotype.** Intra-TAD duplications (blue) do not change the overall TAD conformation but can result in abnormal gene regulation (sex reversal). The centromeric and telomeric parts of the tandem duplications are indicated. Inter-TAD duplications crossing TAD-boundaries (green and yellow) result in the formation of new chromatin domains (neo-TADs). Insulation from their neighbours results in neutralization of regulatory effects and normal phenotype. Incorporation of genes in the neo-TAD provides the duplicated gene with a novel regulatory landscape and can result in gene misexpression (Cooks syndrome).

confined to a TAD (intra-TAD) have no major effect on TAD structure but can result in increased interaction of duplicated regulatory elements with their target gene. In contrast, duplications that cover parts of two TADs and their boundary (inter-TAD) result in the formation of a new TAD that is insulated from its neighbours. We propose that these newly created domains should be called 'neo-TADs'. Genes that become incorporated in a neo-TAD can be activated by its regulatory elements, thereby eliciting pathogenic effects. Our findings also demonstrate that the genomic effects of structural variations cannot solely be explained by the rewiring of enhancer–promoter contacts<sup>3</sup>. Rather, our data show that TADs are robust and stable genomic units that can be rearranged and recombined to create new regulatory regions of the genome. The integrity of these units in relation to neighbouring TADs and genes determines their gene regulatory and thus pathogenic effect. Figure 4 shows a schematic of the proposed disease mechanism associated with TAD changes.

The concept presented here provides a framework to predict the phenotypic outcome of genomic variations that can be directly applied for the interpretation of copy number variations (CNVs) detected in diagnostic screens, routinely performed in patients with congenital malformations and/or intellectual disability<sup>16</sup>, or for structural variations found in cancer<sup>17</sup>. Furthermore, the effects of genomic rearrangements described here are probably important for evolutionary mechanisms, as duplications are thought to be a major driver in evolution<sup>18,19</sup>. The process of gene neofunctionalization is thought to work through gene duplications and subsequent adaptation of one of the gene copies. Our data suggest a further mechanism in which the isolation of a newly formed TAD can result in a phenotypic change in the organism that is then directly subject to selective pressure, without affecting the parent copy of the gene. With variable shifting of TADs and recombination of regulatory activity with new target genes, an entire toolbox of possibilities for new gene functions can be acquired.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 2 February; accepted 23 August 2016.**

**Published online 5 October 2016.**

1. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).

2. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
3. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
4. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24**, 390–400 (2014).
5. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
6. Benko, S. *et al.* Disruption of a long distance regulatory region upstream of *SOX9* in isolated disorders of sex development. *J. Med. Genet.* **48**, 825–830 (2011).
7. Wagner, T. *et al.* Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene *SOX9*. *Cell* **79**, 1111–1120 (1994).
8. Kurth, I. *et al.* Duplications of noncoding elements 5' of *SOX9* are associated with brachydactyly-anonychia. *Nature Genet.* **41**, 862–863 (2009).
9. Gordon, C. T. *et al.* Long-range regulation at the *SOX9* locus in development and disease. *J. Med. Genet.* **46**, 649–656 (2009).
10. Kim, G.-J. *et al.* Copy number variation of two separate regulatory regions upstream of *SOX9* causes isolated 46,XY or 46,XX disorder of sex development. *J. Med. Genet.* **52**, 240–247 (2015).
11. Sekido, R. & Lovell-Badge, R. Sex determination involves synergistic action of SRY and SF1 on a specific *Sox9* enhancer. *Nature* **453**, 930–934 (2008).
12. Dahal, G. R. *et al.* An inwardly rectifying K<sup>+</sup> channel is required for patterning. *Development* **139**, 3653–3664 (2012).
13. Mundlos, S. The brachydactylies: a molecular disease family. *Clin. Genet.* **76**, 123–136 (2009).
14. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
15. Guo, Y. *et al.* CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* **162**, 900–910 (2015).
16. Lupski, J. R. Genomic rearrangements and sporadic disease. *Nature Genet.* **39** (Suppl), S43–S47 (2007).
17. Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005–1010 (2009).
18. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
19. Katju, V. & Berghthorsson, U. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Front. Genet.* **4**, 273 (2013).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We are grateful to all members of the MPIMG transgene and mouse facility for embryonic stem cell aggregation and mouse husbandry. This work was supported by grants from the Deutsche Forschungsgemeinschaft to S.M. and F.S., the BIH to D.M.I., S.M. and A.P., and the Max Planck Foundation to S.M.

**Author Contributions** M.F., F.S. and S.M. conceived the study and designed the experiments. M.F. and G.A. performed 4C-seq, capture Hi-C, with analysis by V.H., D.M.I. and R.S. M.F. and W.S. performed the LacZ staining and analysis. M.F. and D.M.I. performed RNA-seq, *in situ* hybridizations and phenotype analysis. W.-L.C. and I.J. contributed to histological analysis. M.F., W.S., R.K., D.M.I., K.K. and L.W. generated the transgenic mouse models. I.K., P.C., O.Z., G.H., M.S., L.L. and F.B. obtained the patient samples. A.P., W.S., F.S., M.S., M.V. and B.T. contributed to scientific discussion and technical support. M.F., D.M.I. and S.M. wrote the paper with input from all authors.

**Author Information** Sequencing data has been deposited in Gene Expression Omnibus (GEO) under accession number GSE78109. Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M. ([mundlos@molgen.mpg.de](mailto:mundlos@molgen.mpg.de)).

**Reviewer Information** *Nature* thanks B. Ren and the other anonymous reviewer(s) for their contribution to the peer review of this work.



## METHODS

No statistical methods were used to predetermine sample size. There was no randomization of experiments, and investigators were not blinded during experiments and outcome assessment.

**ES cell targeting and transgenic mouse strains.** Embryonic stem (ES) cell culture was performed as described previously<sup>20</sup>. ES and feeder cells were tested for mycoplasma contamination using Mycoalert detection kit (Lonza, catalogue number LT07-118) and Mycoalert assay control set (Lonza, catalogue number LT07-518). A single *LoxP* site upstream of *Kcnj16* (mm9 chr11:110,772,110) was targeted in G4-ES cells (129/Sv × C57BL/6 F1 hybrid) using CRISPR/Cas9 and a single-stranded donor oligonucleotide (ssODN). ES cells were cultured on MEF feeder cells under standard conditions and then cotransfected (FuGene HD, Promega) with 8 µg of pX459 vector (Addgene) carrying the single guide RNA (cenKcnj) and 500 pM of a ssODN (40 bp homologous sequence flanking a *LoxP* site), sequences given in Supplementary Table 1. To delete the 18 kb TAD boundary region (Bor) in G4 ES cells and ES cells from Dup-L homozygous mice, two sgRNAs (cenΔBor, telΔBor) were cloned in pX459 vectors and cotransfected. The intra-TAD duplication was generated in G4 ES cells using two sgRNAs (cenDupS, telDupS). Embryos and live animals from ES cells were generated by diploid or tetraploid complementation<sup>21</sup>. Genotyping was performed by PCR analysis.

The SB-Kcnj and SB-Sox9 alleles contain an insertion of Sleeping Beauty (SB) transgene<sup>22</sup>. This transgene harbours a single *LoxP* site and a *lacZ* reporter gene, flanked by transposons. Both alleles were targeted using standard protocols for homologous recombination in E14 ES cells<sup>23</sup>. Primer sequences for amplifying homology sequences are provided in Supplementary Table 1. Positive ES cell clones were injected into donor blastocysts to generate chimaeras. The SB-Kcnj allele was further used for remobilization of the SB transgene, following the protocol in ref. 22, to generate new SB insertion sites at the locus. Duplications and corresponding deletions were generated as described previously<sup>22</sup>. All generated duplications/deletions and donor mouse strains are listed in Extended Data Fig. 4.

Mouse strains were maintained by crossing them with C57BL6/J mice. All animal procedures were conducted as approved by the local authorities (LAGeSo Berlin) under the license numbers G0368/08 and G0247/13.

**ES cell generation of Dup-L mice.** ES cells from Dup-L homozygous blastocysts were established using N2B27 medium supplemented with FGF/Erk and Gsk3 pathway inhibitors (2i) and LIF according to ref. 24. An established ES cell line was used for subsequent boundary deletion experiments using CRISPR/Cas9.

**4C-seq.** 4C-seq libraries were generated from microdissected mouse tissues or human fibroblasts as described previously<sup>25</sup>. The starting material for all 4C-seq libraries was  $5 \times 10^6$  to  $1 \times 10^7$  cells, which corresponds to limb buds from a pool of eight E12.5 embryos. All 4C-seq experiments were carried out in heterozygous animals. 4-bp cutters were used as primary and secondary restriction enzymes (Supplementary Table 2). For each viewpoint, a total of 1 to 1.6 µg DNA was amplified by PCR (Supplementary Table 2). All samples were sequenced with Illumina Hi-Seq technology according to standard protocols. 4C-seq experiments from all viewpoints were carried out in biological replicates in wild type and Dup-L mutants and tested for reproducibility (Pearson *R* for WT:Sox9 = 0.87, *Kcnj2* = 0.90; Dup-L:Sox9 = 0.94, *Kcnj2* = 0.94, *lacZ* = 0.89). A representative result is shown in the figures. Experiments from patient samples and Dup-C mutants were performed as singletons.

**Bioinformatics.** For 4C-seq data analysis, reads were pre-processed, mapped to a corresponding reference (GRCh37/hg19 or NCBI37/mm9) using BWA<sup>26</sup> and coverage normalized as reported previously<sup>3</sup>. The viewpoint and adjacent fragments 1.5 kb upstream and downstream were removed and a window of 10 fragments was chosen to normalize the data per million mapped reads (RPM). To compare interaction profiles of different samples, we obtained the log<sub>2</sub>-fold change for each window of normalized reads. To obtain ratios duplicated regions were excluded for calculation of the scaling parameter used in RPM normalization. Code is available upon request.

**SNP analysis of 4C-seq data.** To indicate selective interactions in a patient with an inter-TAD duplication (no phenotype), raw reads of 4C-seq experiments with viewpoints in *SOX9*, *KCNJ2* and in the duplication breakpoint were mapped using BWA<sup>26</sup> without pre-processing and variants were called for all samples together using GATK (v3.4-46)<sup>27</sup> for the duplicated region (chr17:68195430–69981335, hg19). Bi-allelic SNP positions with a minimum coverage of 10 reads were selected. These were called homozygous for one allele contacted by the breakpoint viewpoint and homozygous for the other allele contacted by either the *SOX9* or *KCNJ2* viewpoint. Additionally, whole-genome sequencing was performed and alignment and variant calling was done in the same way. Allele frequencies of the WGS experiment were then compared to the allele frequencies of the 4C-seq experiments at the selected SNP positions.

**SureSelect design.** The library of SureSelect enrichment probes were designed over the genomic interval (mm9, chr11:109010000–114878000) using the SureDesign

online tool of Agilent. Probes are covering the entire genomic region and were not designed specifically in proximity of DpnII sites. The probes covered 88% of the interval.

**Capture Hi-C.** cHi-C libraries were prepared from homozygous E12.5 limb buds (except Dup-S) as described previously<sup>25</sup>. In summary,  $5 \times 10^6$  to  $1 \times 10^7$  cells were used for crosslinking, cell lysis, DpnII digestion, ligation and de-crosslinking. Re-ligated products were then sheared using a Covaris sonicator (duty cycle: 10%, intensity: 5, cycles per burst: 200, time: 6 cycles of 60 s each, set mode: frequency sweeping, temperature: 4 to 7 °C). Adaptors were added to the sheared DNA and amplified according to Agilent instructions for Illumina sequencing. The library was hybridized to the custom-designed sure-select beads and indexed for sequencing (50 to 100 bp paired-end) following Agilent instructions. Capture Hi-C experiments were performed as singletons. As an internal control, we compared the results from six experiments for regions outside of the region of interest (chr11:109,010,001–110,250,000 and chr11:113,100,001–114,870,000). The cHi-C maps of the internal control were highly correlated between the six samples (Spearman *R*: WT/Dup-L = 0.96; WT/Dup-C = 0.94; WT/Dup-S = 0.96; WT/ΔBor = 0.96; WT/Dup-LΔBor = 0.96; Dup-L/Dup-C = 0.95; Dup-L/Dup-S = 0.96; Dup-L/ΔBor = 0.96; Dup-L/Dup-LΔBor = 0.97; Dup-C/Dup-S = 0.94; Dup-C/ΔBor = 0.94; Dup-C/Dup-LΔBor = 0.95; Dup-S/ΔBor = 0.96; Dup-S/Dup-LΔBor = 0.96; ΔBor/Dup-LΔBor = 0.96) confirming the high reproducibility of the methodology.

**Bioinformatics.** Preprocessing and mapping of paired-end sequencing data, as well as filtering of mapped di-tags was performed with the HiCUP pipeline v.0.5.8 (ref. 28). The pipeline used Bowtie2 v.2.2.6 (ref. 29) for mapping short reads to reference genome (NCBI37/mm9). Filtered di-tags were further processed with Juicebox<sup>3</sup> command line tools to bin di-tags (10 kb bins) and to normalize the map by KR normalization. For this, only reads with a MAPQ ≥ 30 were considered. The DNA-capturing step enriches genomic region chr11:109,010,001–114,878,000 on mm9 leading to three different regimes in the cHi-C map: (i) enriched versus enriched, (ii) enriched versus non-enriched, and (iii) non-enriched versus non-enriched. For binning and normalization only di-tags in regime (i) were considered. Therefore di-tags were filtered for the enriched region and mm9 coordinates were shifted by –109,010,000 bp. For Juicebox a custom chromosome sizes file containing only the enriched region on chr11 (length 5,868,000 bp) was used. After binning and normalization, coordinates were shifted back to their original values. However, in general duplicated regions yield more signal compared to non-duplicated regions when mapped to the wildtype reference genome. The signal is flattened disproportionately in duplicated regions by a normalization procedure that balances the whole interaction matrix, such as KR normalization. Therefore, we used only raw count maps to calculate the differences between samples. Difference maps were generated based on raw count maps, scaled individually by dividing each value of the matrix by a factor (sum ‘masked’ triangle matrix/10<sup>6</sup>). To avoid copy number biases, the region spanning all tested duplications (chr11: 110,770,001–112,520,000) was not considered for the computation of the scaling factor. cHi-C maps of count values and difference maps were visualized with the WashU epigenome browser<sup>30</sup>.

**RNA-seq.** E12.5 distal limbs were microdissected from wild-type and mutant embryos (*n* = 2) and immediately frozen in liquid nitrogen. From these RNA was isolated using TRIzol extraction and the RNeasy Mini Kit (QIAGEN). Samples were poly-A enriched and sequenced (paired-end 50 bp) using Illumina technology following standard protocols.

**Bioinformatics:** 50 bp paired-end reads were mapped to the mouse reference genome (mm9) using the STAR mapper<sup>31</sup> (splice junctions based on RefSeq; options: –alignIntronMin 20–alignIntronMax 500000–outFilterMultimapNmax 5–outFilterMismatchNmax 10–outFilterMismatchNoverLmax 0.1). Reads per gene were counted as described previously<sup>3</sup>, and used for differential expression analysis with the DESeq2 package<sup>32</sup>.

**Phenotypic analysis.** Phenotypic analysis for mutant mouse lines was carried out for at least three animals per analysis and developmental stage. Phenotypic analysis was carried out in heterozygous animals, for ΔBor in heterozygous and homozygous animals, and for Dup-LΔBor in homozygous animals. Penetrance of the phenotypes was investigated by analysing >30 offspring and considered fully penetrant if all animals were similarly affected.

**Micro-computer tomography.** Autopods of seven-week-old control and mutant mice (*n* = 3) were scanned using a Skyscan 1172 X-ray microtomography system (Brucker microCT, Belgium) at 5 µm resolution. 3D model reconstruction was done with the Skyscan image analysis software CT-Analyser and CT-volume (Brucker microCT, Belgium).

**Whole-mount *in situ* hybridization and histology.** Wild-type and mutant E12.5 embryos (*n* = 5) were subjected to whole mount *in situ* hybridization using standard procedures. *Sox9* and *Kcnj2* probes were generated by PCR amplification using mouse limb bud cDNA (Supplementary Table 1).

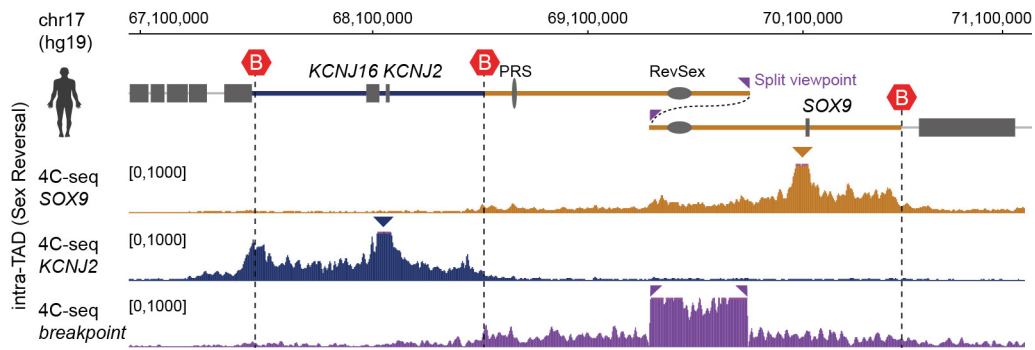
For Masson–Goldner staining, limbs of newborns (P0,  $n=2$ ) were first fixed in 4% PFA in PBS containing 0.5 M EDTA and then embedded in paraffin. Paraffin sections of 6  $\mu$ m were then stained with the Merck Masson–Goldner staining kit (catalogue number 100485) and Meyers haematoxylin (catalogue number MHS80) according to the manufacturer's instructions.

**LacZ staining and optical projection tomography.** E12.5 mouse embryos ( $n=3$ ) were dissected in cold PBS, fixed in 4% PFA in PBS on ice for 30 min, washed twice with ice-cold PBS and once at room temperature (19–24 °C), and then stained overnight for  $\beta$ -galactosidase activity in a humid chamber at 37 °C as previously described<sup>22</sup>. After staining, embryos were washed in PBS and stored at 4 °C in 4% PFA in PBS.

For OPT scanning, stained embryos were embedded in 1% low-melt agarose and dehydrated over 1–2 days using methanol (2–3 methanol steps). Subsequently samples were cleared overnight in BABB (1 part benzyl alcohol, 2 parts benzyl benzoate). The samples were then scanned (802 frames) with a Biotronics OPT 3001 M scanner with white light (exposure = 10–100 ms) for LacZ staining using GFP filter combination (exciter 425 nm/40 nm - emitter LP475 nm) and 20–380 ms exposure for autofluorescence. 3D reconstructions were generated using Skyscan software and further analysed with Amira and Imaris software.

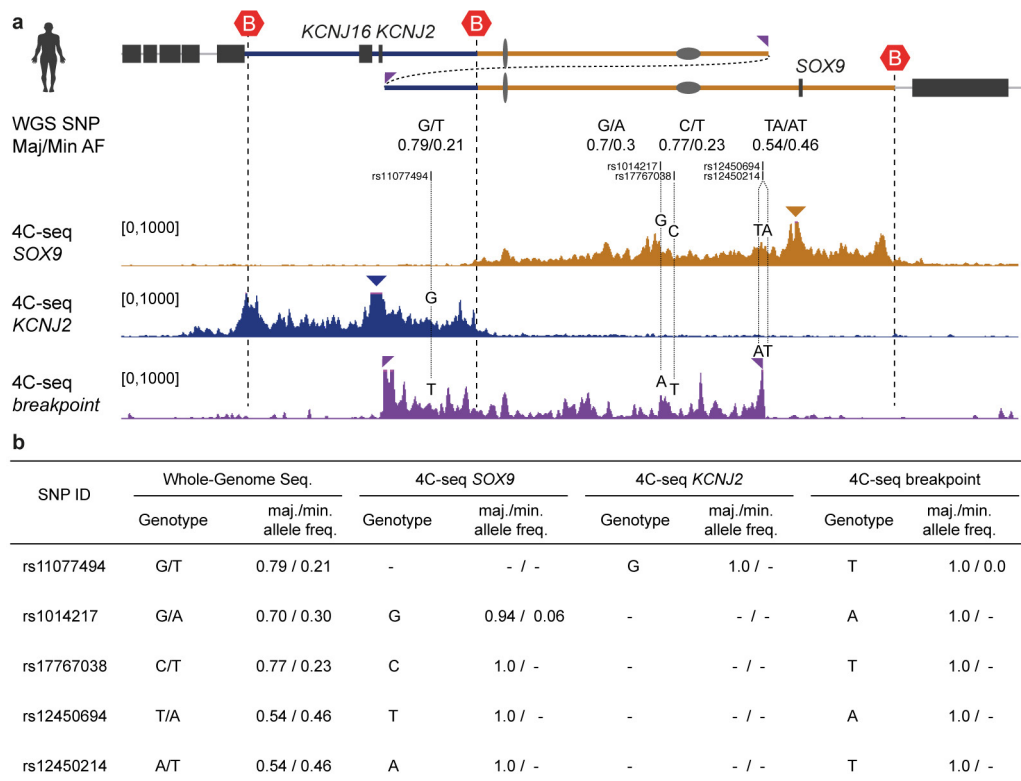
**Human material.** Skin biopsies from one sex reversal patient, the no phenotype individual, one Cooks syndrome patient and controls, as well as gonadal ('testicular') fibroblasts from one sex reversal individual were obtained by standard procedures. Fibroblasts were cultured in DMEM (Lonza) supplemented with 10% fetal calf serum (Gibco), 1% L-glutamine (Lonza) and 1% penicillin/streptomycin (Lonza). Written informed consent was obtained from all individuals studied that participated in this study. This study was approved by the Charité Universitätsmedizin Berlin ethics committee.

20. Kraft, K. *et al.* Deletions, inversions, duplications: engineering of structural variants using CRISPR/Cas in Mice. *Cell Reports* **10**, 833–839 (2015).
21. Artus, J. & Hadjantonakis, A.-K. in *Transgenic Mouse Methods and Protocols* Vol. 693 *Methods in Molecular Biology* (eds Hofker, M. H. & van Deursen, J.) Ch. 3, 37–56 (Humana Press, 2011).
22. Ruf, S. *et al.* Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nature Genet.* **43**, 379–386 (2011).
23. Hooper, M., Hardy, K., Handyside, A., Hunter, S. & Monk, M. HPRT-deficient (Lesch–Nyhan) mouse embryos derived from germline colonization by cultured cells. *Nature* **326**, 292–295 (1987).
24. Nagy, K. & Nichols, J. in *Advanced Protocols for Animal Transgenesis* (eds Pease, S. & Saunders, T. L.) Ch. 18, 431–455 (Springer, 2011).
25. van de Werken, H. J. G. *et al.* in *Methods in Enzymology* Vol. 513 (eds Wu, C. & Allis, D. C.) 89–112 (Academic Press, 2012).
26. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
27. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
28. Wingett, S. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res.* **4**, 1310 (2015).
29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
30. Zhou, X. *et al.* Exploring long-range genome interactions using the WashU Epigenome Browser. *Nature Methods* **10**, 375–376 (2013).
31. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
32. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
33. Bhatia, S. *et al.* Functional assessment of disease-associated regulatory variants *in vivo* using a versatile dual colour transgenesis strategy in zebrafish. *PLoS Genet.* **11**, e1005193 (2015).
34. Benko, S. *et al.* Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nature Genet.* **41**, 359–364 (2009).
35. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
36. Gordon, C. T. *et al.* Identification of novel craniofacial regulatory domains located far upstream of SOX9 and disrupted in Pierre Robin sequence. *Hum. Mutat.* **35**, 1011–1020 (2014).
37. Yao, B. *et al.* The SOX9 upstream region prone to chromosomal aberrations causing campomelic dysplasia contains multiple cartilage enhancers. *Nucleic Acids Res.* **43**, 5394–5408 (2015).



**Extended Data Figure 1 | 4C-seq from patient fibroblasts with a 470 kb intra-TAD duplication (sex reversal).** Schematic representation depicts TAD structure with *KCNJ* TAD (blue), *SOX9* TAD (brown) and TAD boundaries (red hexagons). Size and position of the 470 kb intra-TAD duplication in a patient with female-to-male sex reversal is indicated by the overlap. Patient fibroblasts were derived from a gonadal biopsy of ovo-testes. 4C-seq tracks below show interaction profiles from *SOX9*

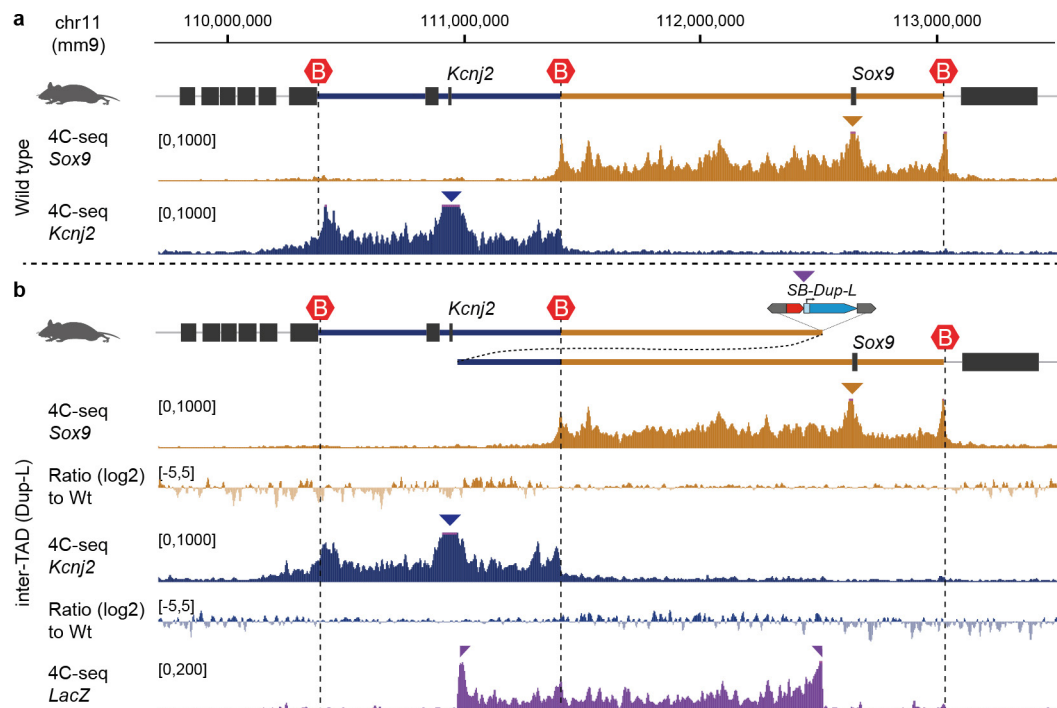
(brown), *KCNJ2* (blue) and breakpoint (purple) viewpoint. Note interaction profiles from *SOX9* and *KCNJ2* are restricted to their corresponding TADs. Unique viewpoint at the breakpoint shows that interactions are restricted to the *SOX9* TAD. Note all reads mapped to a wild-type genome resulting in split viewpoint for the duplication breakpoint.



**Extended Data Figure 2 | Allele frequencies determined in 4C-seq data indicate selective interactions in a patient with an inter-TAD duplication (no phenotype).** SNP analysis related to Fig. 1d. **a**, Schematic of SNP-analysis in a patient carrying an inter-TAD duplication (heterozygous) with no phenotype. SNP positions and their allele frequency (AF) identified by whole-genome sequencing from the patient are shown. Bottom, 4C-seq interaction profiles using

SOX9 (brown), KCNJ2 (blue) and the duplication breakpoint (purple) as viewpoints (triangles). SNPs contacted by individual viewpoints are indicated. Note that variants contacted by the breakpoint viewpoint are not contacted by SOX9 and KCNJ2 viewpoints, suggesting the formation of an insulated interaction domain (purple). **b**, Summary of observed SNP genotype and allele frequency from whole-genome sequencing data and 4C-seq viewpoints.



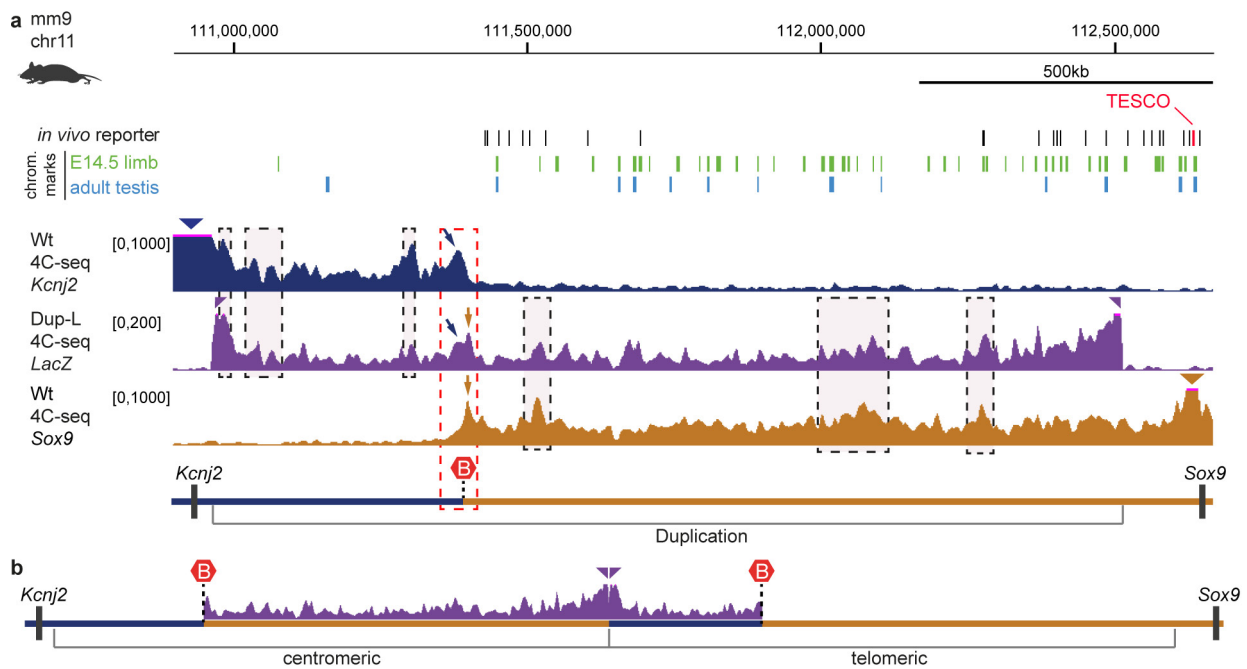


**Extended Data Figure 3 | 4C-seq from an inter-TAD duplication in Dup-L mouse mutants. a,** Schematic representation depicts TAD structure in wild type with *Kcnj2* TAD (blue), *Sox9* TAD (brown) and TAD boundaries (red hexagons). Below, 4C-seq interaction profiles of viewpoints (triangles) in *Sox9* (brown) and *Kcnj2* (blue) from E12.5 wild-type limb buds. **b,** Schematic representation of Dup-L allele. Position of *lacZ* reporter at the duplication breakpoint is shown and duplication is

indicated by overlap. Note that 4C-seq reads are mapped to the wild-type genome, which results in split viewpoint from *lacZ* viewpoint. 4C-seq profiles with viewpoint in *Sox9* (brown), *Kcnj2* (blue) and *lacZ* reporter (purple) in Dup-L are shown below. *Kcnj2* and *Sox9* profiles are unchanged, whereas the unique viewpoint in the *lacZ* reporter shows interactions that are restricted to the duplicated region, suggesting formation of a separate interaction domain.

Mouse alleles			
Name / genomic location	rearrangement / linear locus		
Wild type		Observed phenotype in het. mice	Human equivalent
SB-Kcnj		no abnormal phenotype observed	not applicable
genomic position of SB-insertion: chr11:110,959,589			
SB-Sox9		no abnormal phenotype observed	not applicable
genomic position of SB-insertion: chr11:112,514,692			
Dup-L (inter-TAD)		no abnormal phenotype observed	no abnormal phenotype observed
deleted region: chr11:110,959,589-112,514,692 (1,56 Mb)			
Del-L (inter-TAD)		embryonic lethal, acampomelic dysplasia	campomelic dysplasia
deleted region: chr11:110,955,589-112,514,692 (1,56 Mb)			
Dup-C (inter-TAD)		Cooks syndrome-like	Cooks syndrome
deleted region: chr11:110,772,110-112,514,692 (1,74 Mb)			
Dup-C <sub>2</sub> (inter-TAD)		Cooks syndrome-like	Cooks syndrome
deleted region: chr11:110,838,025-112,514,692 (1,68 Mb)			
Dup-S (intra-TAD)		no abnormal phenotype observed	female-to-male sex reversal
deleted region: chr11:111,752,617-112,172,936 (420 kb)			
Dup-K <sub>1</sub> (intra-TAD)		no abnormal phenotype observed	not known
deleted region: chr11:110,772,110-111,373,724 (601 kb)			
Dup-K <sub>2</sub> (intra-TAD)		no abnormal phenotype observed	not known
deleted region: chr11:110,772,110-110,959,589 (187 kb)			
ΔBor		no abnormal phenotype observed	not known
Boundary deletion: chr11:111,383,859-111,402,200 (18,3 kb)			
Dup-LΔBor		embryonic lethal (ESC quality)	not known
deleted region: chr11:110,959,589-112,514,692 (1,55 Mb)			
Boundary deletion: chr11:111,383,859-111,402,200 (18,3 kb)			

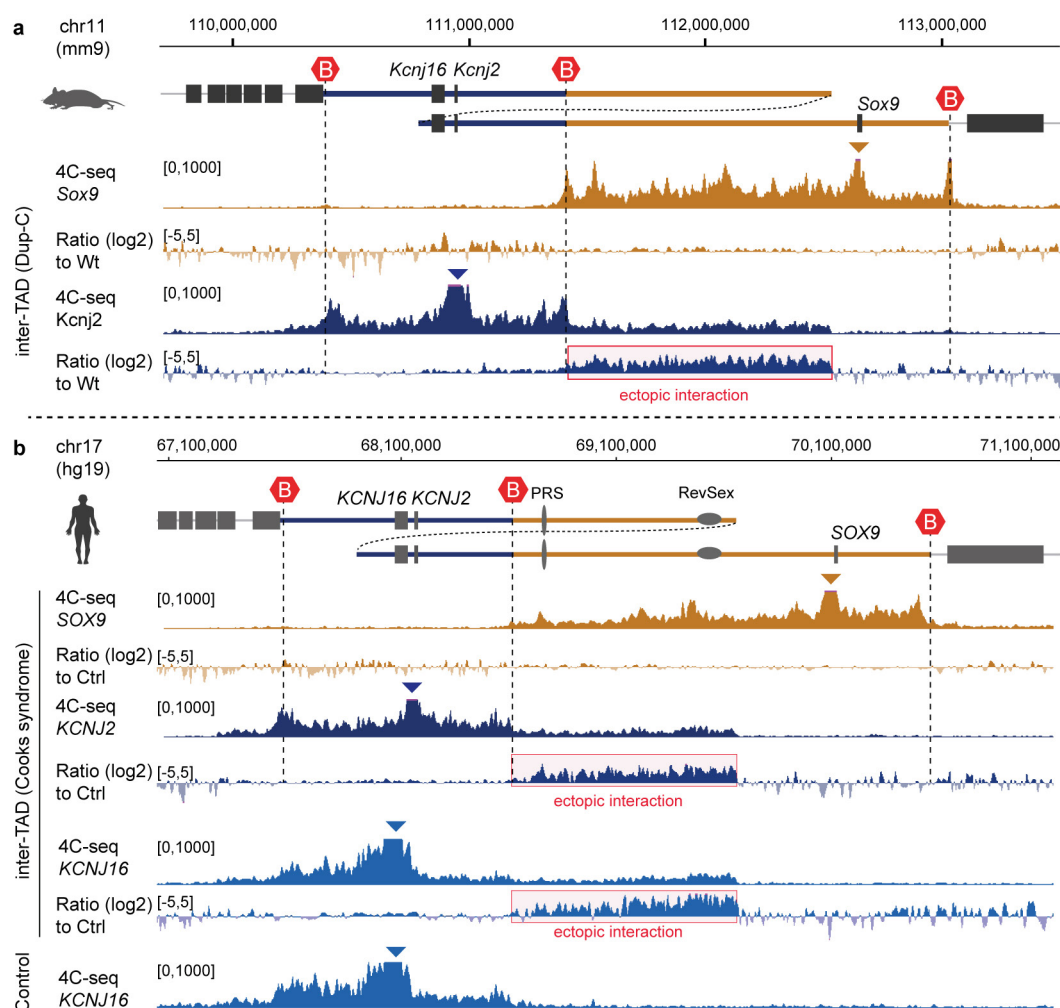
**Extended Data Figure 4 | Overview of mouse alleles used in this study.** The extent of each duplicated and deleted region in mice are given as mm9 coordinates. Observed mouse phenotypes and human syndromes with the equivalent mutation are listed. For generation of alleles see Methods. SB, Sleeping Beauty transgene containing a single *loxP* site (red) and a *lacZ* reporter gene (blue) flanked by SB-transposons.



**Extended Data Figure 5 | 4C-seq interaction profile of neo-TAD in Dup-L mutants resembles profiles from *Sox9* and *Kcnj2* viewpoints.**

**a**, The position of published *in vivo* tested reporter constructs<sup>11,33–37</sup> and active chromatin marks (H3K4me1/H3K27ac positive) in *Sox9*-expressing tissue (mouse ENCODE project) are indicated above the 4C tracks. The only known enhancer driving testis-specific expression, TESCO, is indicated. 4C-seq interaction profiles from *Kcnj2* (blue) and *Sox9* (brown) in wild type and from the *lacZ* reporter gene (purple) in E12.5 Dup-L mutants are shown. Triangles indicate viewpoints. Note that 4C-seq reads are mapped to the wild-type genome resulting in split viewpoint from

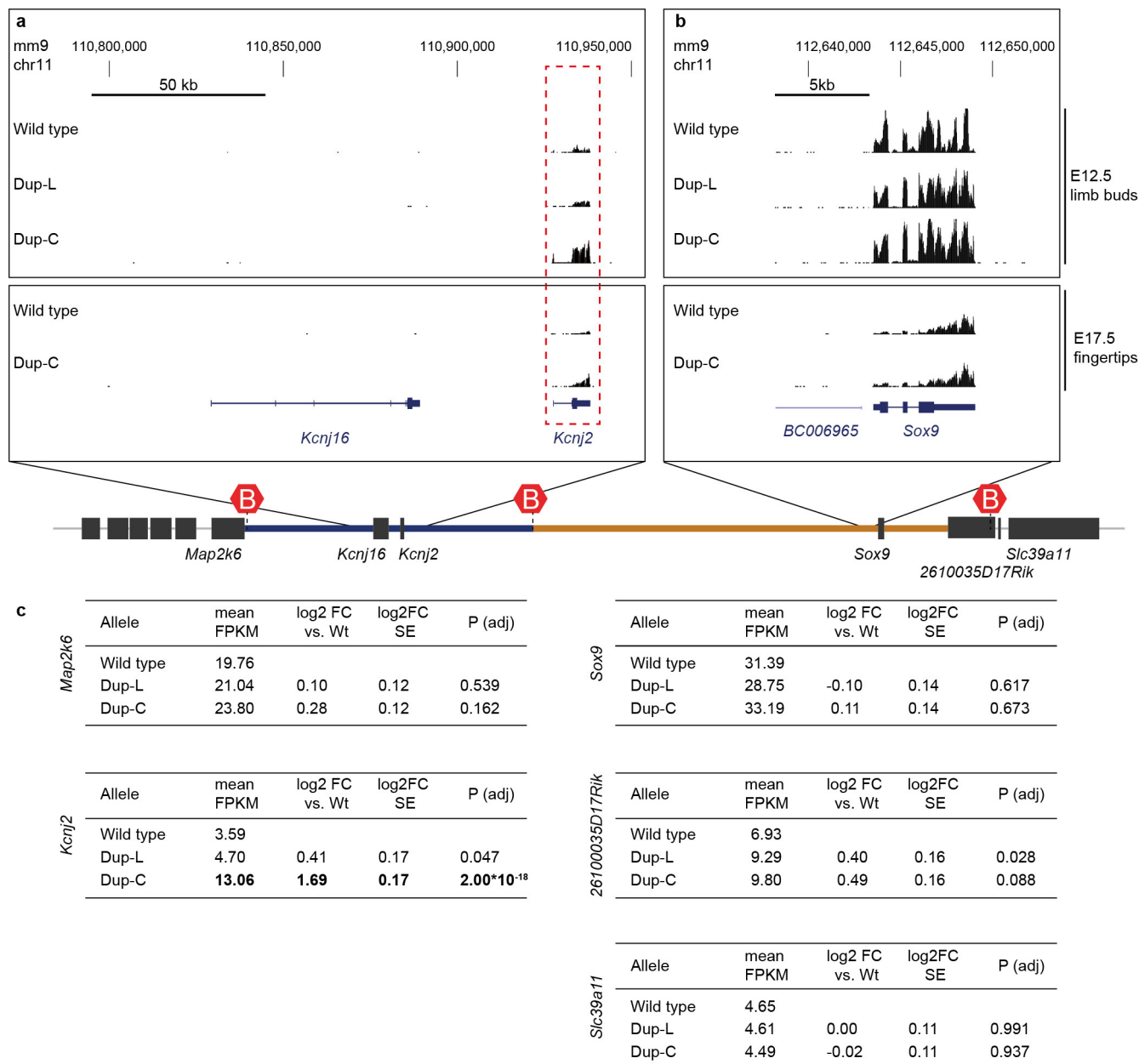
the *lacZ* viewpoint. The interaction profile from *lacZ* viewpoint located at the breakpoint of the Dup-L duplication shows a similar pattern and peak distribution (dashed boxes) as the *Sox9* and *Kcnj2* viewpoints. Note similarities of peak profiles at the boundary between *Kcnj2* and *Sox9* TADs (arrows in red dashed box). **b**, Schematic of locus with an artificially duplicated region and corresponding 4C-seq profile from the *lacZ* viewpoint inside the neo-TAD in Dup-L mutants. The centromeric and telomeric part of the tandem duplication is indicated. The purple 4C-seq profile corresponds to the neo-TAD delimited by the duplicated TAD boundary.



**Extended Data Figure 6 | 4C-seq of inter-TAD Cooks syndrome duplications in mouse and human.** Extent and position of the duplications is indicated by the overlap in the schematic. 4C-seq profiles with indicated viewpoints and ratio to control (below) **a**, *Sox9* (brown) and *Kcnj2* (blue) 4C-seq from Dup-C mutant limb buds at E12.5. **b**, 4C-seq from fibroblasts of an individual with Cooks syndrome and inter-

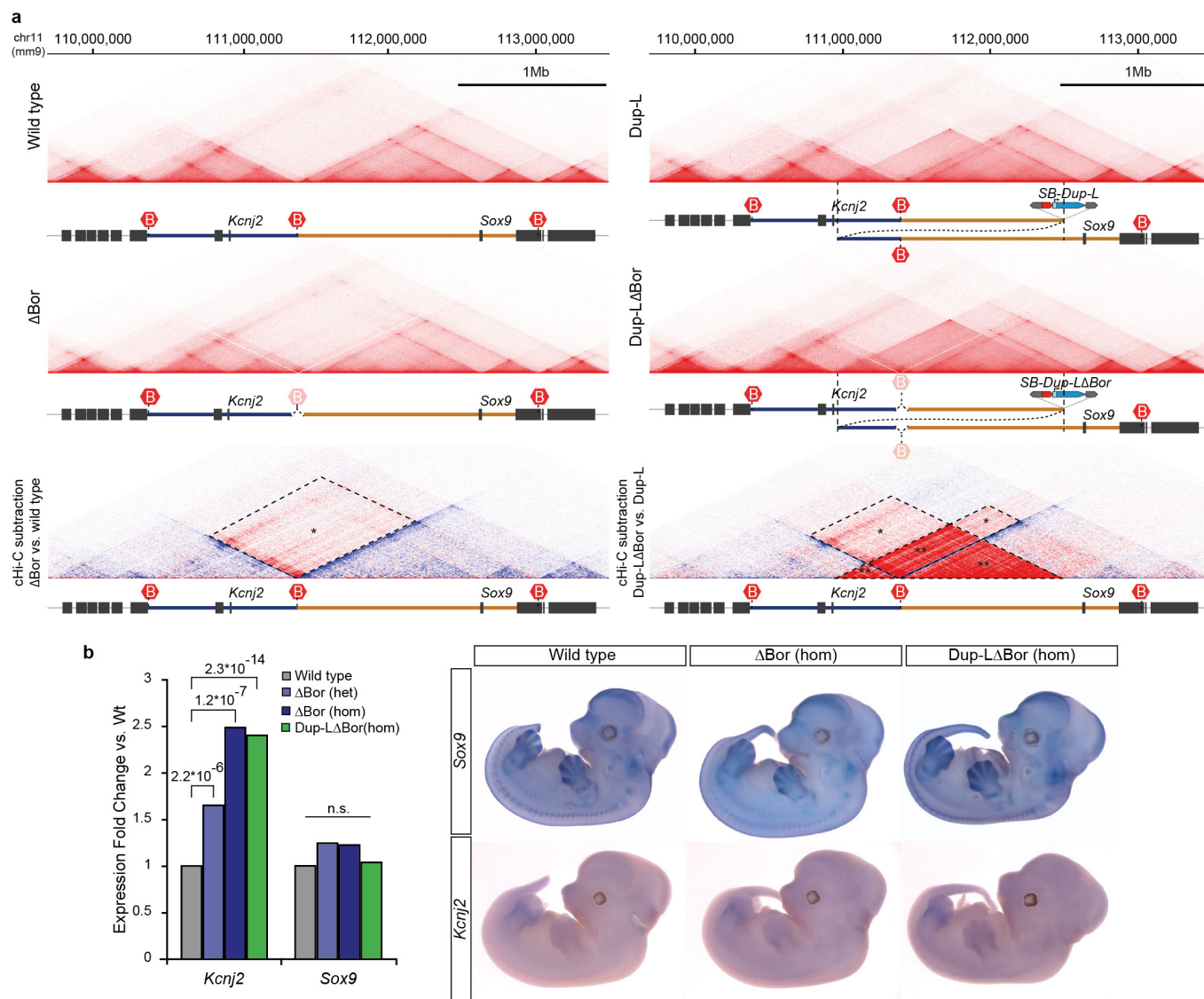
TAD duplication from *SOX9* and *KCNJ2* viewpoints. Below, 4C-seq with viewpoint in *KCNJ16* (light blue) in patient and control fibroblasts and the ratio of patient to control. **a**, **b**, Incorporation of *Kcnj/KCNJ* genes in the neo-TAD results in ectopic contacts from the *Kcnj2* or *KCNJ2* and *KCNJ16* viewpoints with the duplicated part of the *Sox9* TAD, whereas interactions from the *Sox9/SOX9* viewpoint remain unchanged.





**Extended Data Figure 7 | RNA-seq analysis of genes at the *Sox9*, *Kcnj2* and *Kcnj16* locus from mouse mutants used in this study. a, b, RNA-seq from wild-type and mutant E12.5 limb buds and E17.5 fingertips (lower two tracks) shows read profiles for *Kcnj16*, *Kcnj2* and *Sox9*. Note the absence of expression of *Kcnj16* in the examined tissue. c, Summary**

of expression values of genes at the *Sox9*, *Kcnj2* and *Kcnj16* locus from E12.5 limb buds. Significant expression changes are in bold. FPKM, fragments per kilobase of exon per million fragments mapped. Benjamini-Hochberg-adjusted *P* value, *n* = 2, cut-off = 0.001.



**Extended Data Figure 8 | Deletion of the TAD boundary region in wild-type and Dup-L mutant mice. a**, Left, cHi-C from wild type (top) and  $\Delta$ Bor (middle) E12.5 limb buds and subtraction map of  $\Delta$ Bor relative to wild type (bottom). Deletion of the boundary between the *Kcnj2* and *Sox9* TADs leads to loss of insulation and ectopic interactions (\*) between the TADs. Note overall TAD structure remains intact and ectopic interaction is restricted by remaining TAD boundaries. Right, cHi-C from Dup-L (top) and Dup-L $\Delta$ Bor (middle) E12.5 limb buds and subtraction map of Dup-L $\Delta$ Bor relative to Dup-L (bottom). Deletion of the two duplicated boundary regions flanking the neo-TAD, results in ectopic contacts of

duplicated sequences with adjacent *Kcnj2* and *Sox9* TADs (\*) including *Sox9* and *Kcnj2*, as seen in  $\Delta$ Bor mutants. Loss of neo-TAD insulation results in increased interactions of the duplicated sequences (\*\*).

**b**, Expression analysis of *Sox9* and *Kcnj2* in wild-type,  $\Delta$ Bor and Dup-L $\Delta$ Bor embryos. RNA-seq expression analysis of mutant versus wild-type E12.5 limb buds. *Kcnj2* is upregulated in  $\Delta$ Bor (heterozygous and homozygous) and Dup-L $\Delta$ Bor (homozygous) limb buds (Benjamini-Hochberg adjusted *P* values, *n* = 2). Whole mount *in situ* hybridization shows no site-specific misexpression of *Kcnj2* in a *Sox9*-like pattern.

# Two distinct RNase activities of CRISPR–C2c2 enable guide–RNA processing and RNA detection

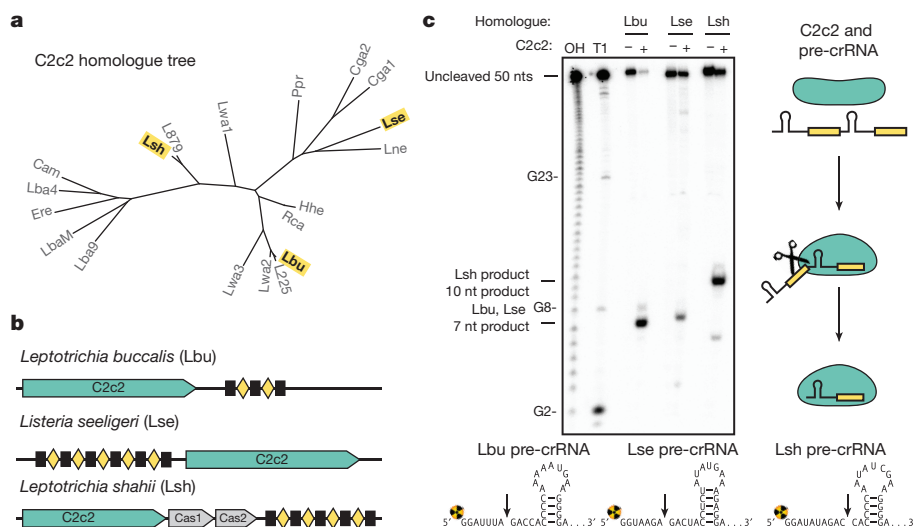
Alexandra East-Seletsky<sup>1\*</sup>, Mitchell R. O'Connell<sup>1\*</sup>, Spencer C. Knight<sup>2</sup>, David Burstein<sup>3</sup>, Jamie H. D. Cate<sup>1,2,4</sup>, Robert Tjian<sup>1,5,6,7</sup> & Jennifer A. Doudna<sup>1,2,4,6,8</sup>

Bacterial adaptive immune systems use CRISPRs (clustered regularly interspaced short palindromic repeats) and CRISPR-associated (Cas) proteins for RNA-guided nucleic acid cleavage<sup>1,2</sup>. Although most prokaryotic adaptive immune systems generally target DNA substrates<sup>3–5</sup>, type III and VI CRISPR systems direct interference complexes against single-stranded RNA substrates<sup>6–9</sup>. In type VI systems, the single-subunit C2c2 protein functions as an RNA-guided RNA endonuclease (RNase)<sup>9,10</sup>. How this enzyme acquires mature CRISPR RNAs (crRNAs) that are essential for immune surveillance and how it carries out crRNA-mediated RNA cleavage remain unclear. Here we show that bacterial C2c2 possesses a unique RNase activity responsible for CRISPR RNA maturation that is distinct from its RNA-activated single-stranded RNA degradation activity. These dual RNase functions are chemically and mechanistically different from each other and from the crRNA-processing behaviour of the evolutionarily unrelated CRISPR enzyme Cpf1 (ref. 11). The two RNase activities of C2c2 enable multiplexed processing and loading of guide RNAs that in turn allow sensitive detection of cellular transcripts.

The first step of CRISPR immune surveillance requires the processing of precursor crRNA transcripts (pre-crRNAs), consisting of repeat sequences flanking viral spacer sequences, into individual

mature crRNAs that each contain a single spacer<sup>12–14</sup>. CRISPR systems use three known mechanisms to produce mature crRNAs: a dedicated endonuclease (for example, Cas6 or Cas5d in type I and III systems)<sup>15–17</sup>, coupling of a host endonuclease (for example, RNase III with a trans-activating crRNA in type II systems)<sup>18</sup>, or RNase activity intrinsic to the effector enzyme itself (for example, Cpf1 in type V systems)<sup>11</sup>.

Because type VI CRISPR loci lack an obvious Cas6- or Cas5d-like endonuclease or trans-activating crRNA<sup>10</sup>, we wondered whether C2c2 itself might possess pre-crRNA processing activity, and if so, whether the mechanism would be distinct from Cpf1, an unrelated class 2 CRISPR effector that can process pre-crRNAs<sup>11</sup>. Using purified recombinant C2c2 protein homologues from three distinct branches of the C2c2 protein family (Fig. 1a, b, Extended Data Figs 1, 2), we found that all three enzymes cleave 5'-end radiolabelled pre-crRNA substrates consisting of a full-length consensus repeat sequence and a 20-nucleotide spacer sequence (Fig. 1c). Mapping the cleavage site for each pre-crRNA–C2c2 homologue pair revealed that processing occurs either two or five nucleotides upstream of the predicted repeat-sequence hairpin structure, depending on the C2c2 homologue (Fig. 1c, Extended Data Fig. 3a). Notably, these mapped 5'-cleavage sites do not agree with the



**Figure 1 | C2c2 proteins process precursor crRNA transcripts to generate mature crRNAs.** **a**, Maximum-likelihood phylogenetic tree of C2c2 proteins. Homologues used in this study are highlighted in yellow. **b**, Diagram of the type VI CRISPR loci used in this study. Black rectangles denote repeat elements, yellow diamonds denote spacer sequences. Cas1 and Cas2 are only found in the genomic vicinity of LshC2c2. **c**, C2c2-

mediated cleavage of pre-crRNA derived from the LbuC2c2, LseC2c2 and LshC2c2 CRISPR repeat loci. OH, alkaline hydrolysis ladder; T1, RNase T1 hydrolysis ladder. Processing cleavage reactions were performed with 100 nM C2c2 and <1 nM pre-crRNA. Schematic of cleavage is depicted on the right and predicted pre-crRNA secondary structures are shown below, with arrows indicating the mapped C2c2 cleavage sites (nt, nucleotides).

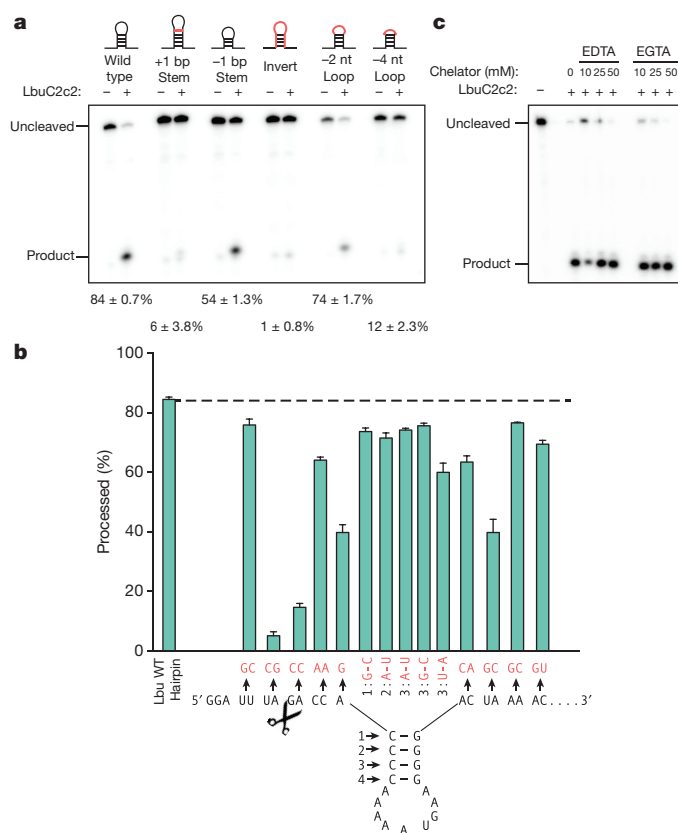
<sup>1</sup>Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. <sup>2</sup>Department of Chemistry, University of California, Berkeley, California 94720, USA.

<sup>3</sup>Department of Earth And Planetary Sciences, University of California, Berkeley, California 94720, USA. <sup>4</sup>MBIB Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

<sup>5</sup>Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147, USA. <sup>6</sup>Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. <sup>7</sup>Li Ka

Shing Biomedical and Health Sciences Center, University of California, Berkeley, California 94720, USA. <sup>8</sup>Innovative Genomics Initiative, University of California, Berkeley, California 94720, USA.

\*These authors contributed equally to this work.



**Figure 2 | LbuC2c2-mediated crRNA biogenesis depends on both structure and sequence of CRISPR repeats.** **a**, Representative cleavage assay by LbuC2c2 on pre-crRNAs containing structural mutations within the stem and loop regions of hairpin. Processed percentages listed below are quantified at 1 h (mean  $\pm$  s.d.,  $n = 3$ ). **b**, Bar graph showing the dependence of pre-crRNA processing on the CRISPR repeat sequence. The wild-type (WT) repeat sequence is shown below with individual bars representing tandem nucleotide mutations as noted in red. The cleavage site is indicated by cartoon scissors. Percentage processed was measured after 1 h (mean  $\pm$  s.d.,  $n = 3$ ). **c**, Divalent metal ion dependence of the crRNA processing reaction was tested by the addition of 10–50 mM EDTA and EGTA to standard reaction conditions.

previously reported sites for *Leptotrichia shahii* (LshC2c2) or *Listeria seeligeri* (LseC2c2) pre-crRNAs<sup>10</sup>. Our analysis of the RNA sequencing dataset in ref. 10 indicates agreement of the *in vivo* cleavage site with the *in vitro* site reported here (Extended Data Fig. 3b–i). Furthermore, cleavage assays using C2c2 from *Leptotrichia buccalis* (LbuC2c2) and a larger pre-crRNA comprising a tandem hairpin-repeat array yielded two products resulting from two separate cleavage events (Extended Data Fig. 4a), consistent with a role for C2c2 in processing precursor crRNA transcripts generated from type VI CRISPR loci.

To understand the substrate requirements and mechanism of C2c2 guide-RNA processing, we generated pre-crRNAs containing mutations in either the stem–loop or the single-stranded flanking regions of the consensus repeat sequence, and tested their ability to be processed by LbuC2c2 (Fig. 2). C2c2-catalysed cleavage was attenuated upon altering the length of the stem in the repeat region (Fig. 2a). Inversion of the stem–loop or reduction of the loop length also reduced the processing activity of C2c2, while contiguous 4-nucleotide mutations including or near the scissile bond completely abolished it (Extended Data Fig. 4b). More extensive mutational analysis of the full crRNA repeat sequence revealed two distinct regions on either side of the hairpin with marked sensitivity to base changes (Fig. 2b). By contrast, there was no dependence on the spacer sequence for kinetics of processing (Extended Data Fig. 4c). This sensitivity to both flanking regions of the

hairpin is reminiscent of the sequence and structural motifs required by many Cas6 and Cas5d enzymes<sup>12,13,19</sup>. Cpf1, however, does not have any dependence on the 3' hairpin flanking region, as the variable spacer region abuts the hairpin stem<sup>11</sup>.

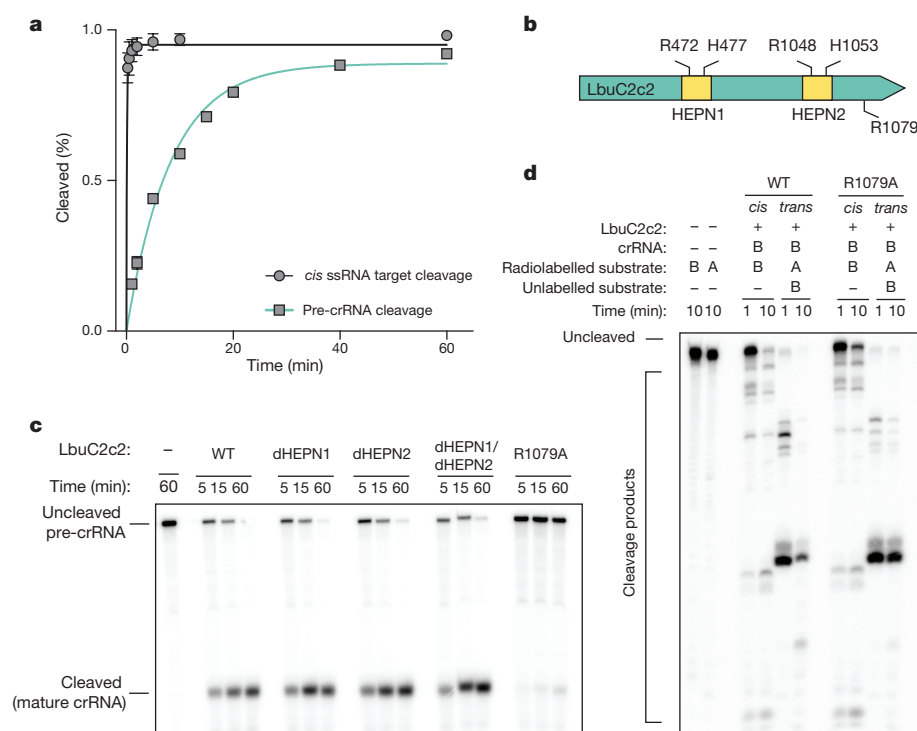
The processing activity of LbuC2c2 was unaffected by the presence of divalent metal ion chelators (Fig. 2c), indicative of a metal-ion-independent RNA hydrolytic mechanism. Metal-ion-independent RNA hydrolysis is typified by the formation of a 2',3'-cyclic phosphate and a 5'-hydroxide on the 5' and 3' halves of the crRNA cleavage products, respectively<sup>20</sup>. To determine the end-group chemical identity of C2c2-processed substrates, we incubated the 5' flanking products with T4 polynucleotide kinase, which removes 2',3'-cyclic phosphates to leave a 3'-hydroxyl; this resulted in altered denaturing-gel migration of the 5' flanking product, consistent with the removal of a 3' phosphate group (Extended Data Fig. 4d). The divalent metal ion independence of the C2c2 pre-crRNA processing activity is in stark contrast with the divalent metal ion dependency of Cpf1, the only other single-protein CRISPR effector shown to perform guide processing<sup>11</sup>. Collectively, these data indicate that C2c2-catalysed pre-crRNA cleavage is a divalent metal-ion-independent process that probably uses a general acid–base catalysis mechanism<sup>20</sup>.

After maturation, crRNAs typically bind with high affinity to Cas effector protein(s) to create RNA-guided surveillance complexes capable of sequence-specific nucleic acid recognition<sup>1,2,21</sup>. In agreement with previous work on LshC2c2 (ref. 9), LbuC2c2 catalysed efficient target RNA cleavage only when such substrates could base pair with a complementary sequence in the crRNA (Extended Data Figs 5–7). Given the promiscuous pattern of cleavage observed for C2c2 (Extended Data Fig. 6), we tested the ability of LbuC2c2 to act as a crRNA-activated non-specific RNA endonuclease *in trans* (Extended Data Fig. 5b). In marked contrast to non-target cleavage experiments performed in *cis* and consistent with observations for LshC2c2 (ref. 9), we observed rapid degradation of non-target RNA (Extended Data Fig. 5b); thus, target recognition activates C2c2 for general non-specific degradation of RNA. Importantly, the similar cleavage rates and near-identical cleavage products observed for both *cis* on-target cleavage and *trans* non-target cleavage of the same RNA substrate implicate the same nuclease centre in both activities (Extended Data Fig. 5b).

crRNA-mediated cleavage of target single-stranded RNA (ssRNA) occurs ~80-fold faster than pre-crRNA processing (Fig. 3a). In contrast to pre-crRNA processing, RNA-guided target cleavage is abolished in the presence of EDTA, indicating that this activity is divalent metal ion-dependent (Fig. 3a, Extended Data Figs 5c, 7). Given these clear differences, we reasoned that C2c2 might possess two orthogonal RNA cleavage activities: crRNA maturation, and crRNA-directed, non-specific RNA degradation. To test this hypothesis, we generated mutants within several residues within the conserved HEPN (higher eukaryotes and prokaryotes nucleotide-binding domain) motifs of LbuC2c2 (refs 9, 22–24) and assessed their pre-crRNA processing and RNA-guided RNase activities (Fig. 3, Extended Data Fig. 7d). Double and quadruple mutants (R472A, H477A, R1048A and H1053A) retained robust pre-crRNA cleavage activity (Fig. 3c), but all HEPN mutations abolished RNA-guided cleavage activity without affecting crRNA or ssRNA binding ability<sup>9</sup> (Extended Data Figs 7d, 8).

We sought mutations that would abrogate pre-crRNA processing activity without disrupting target RNA cleavage. Given that we were unable to predict any other potential RNase motifs beyond the HEPN motifs, and that C2c2 proteins lack homology to Cpf1, we opted to mutate charged residues throughout LbuC2c2 systematically. We identified an arginine residue (R1079A) that upon mutation resulted in severely attenuated pre-crRNA processing activity (Fig. 3c). This C2c2 mutant enzyme retained crRNA-binding ability as well as RNA target cleavage activity (Extended Data Fig. 8d, Fig. 3d). These results show that distinct active sites within the C2c2 protein catalyse pre-crRNA processing and RNA-directed RNA cleavage.





**Figure 3 | LbuC2c2 contains two distinct RNase activities.** **a**, Quantified time-course data of *cis* ssRNA target (black) and pre-crRNA (teal) cleavage by LbuC2c2. Exponential fits are shown as solid lines ( $n = 3$ ), and the calculated pseudo-first-order rate constants ( $k_{\text{obs}}$ ) (mean  $\pm$  s.d.) are  $9.74 \pm 1.15 \text{ min}^{-1}$  and  $0.12 \pm 0.02 \text{ min}^{-1}$  for *cis* ssRNA target and pre-crRNA cleavage, respectively. **b**, LbuC2c2 architecture depicting the location of HEPN motifs and processing-deficient point mutant **c**, **d**, Representative ribonuclease activity of LbuC2c2 mutants for pre-crRNA processing in **c** and ssRNA targeting in **d** and Extended Data Fig. 6d.

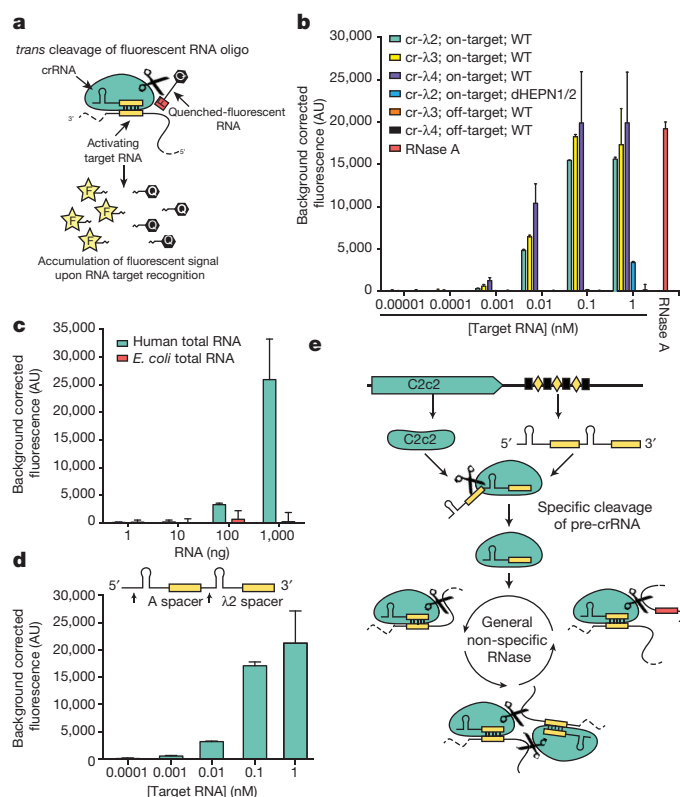
We recognized that the robust RNA-stimulated cleavage of substrates by C2c2 in *trans* might be used to detect specific RNAs within a pool of transcripts. While many polymerase-based methods have been developed for RNA amplification and subsequent detection, few approaches are able to detect target RNA directly without considerable engineering or stringent design constraints for each new RNA target<sup>20,25–27</sup>. As a readily programmable alternative, we tested whether the RNA-guided *trans*-endonuclease activity of C2c2 could be harnessed to cleave a fluorophore quencher-labelled reporter RNA, thereby resulting in increased fluorescence upon target-RNA-triggered RNase activation (Fig. 4a). LbuC2c2 was loaded with bacteriophage  $\lambda$ -targeting crRNAs and tested for its ability to detect the corresponding  $\lambda$  ssRNA targets spiked into HeLa cell total RNA. Upon addition of as little as 1–10 pM complementary  $\lambda$  target RNA, a substantial crRNA-specific fluorescence increase occurred within 30 min (Fig. 4b, Extended Data Fig. 9a). Control experiments with the C2c2–crRNA complex either alone or in the presence of crRNA and a non-complementary target RNA resulted in negligible increases in fluorescence relative to a positive control (Fig. 4b, Extended Data Fig. 9a). At 10 pM concentration of a  $\lambda$  target RNA, only  $\sim 0.02\%$  of the C2c2–crRNA complex is predicted to be in the active state, yet the observed fluorescent signal reflected  $\sim 25$ – $50\%$  cleavage of the reporter RNA substrate, depending on the RNA target. Fragment size resolution of the background RNA in these reactions revealed considerable degradation, even on highly structured tRNAs (Extended Data Fig. 9b). Since reporter RNA cleavage occurs in the presence of a vast excess of unlabelled RNA, we conclude that LbuC2c2 is a robust multiple-turnover enzyme capable of at least  $10^4$  turnovers per target RNA recognized. Thus, in contrast to previous observations<sup>9</sup>, crRNA-directed *trans* cleavage is potent and detectable even at extremely low levels of activated protein.

To extend this LbuC2c2 RNA detection system, we designed a crRNA to target endogenous  $\beta$ -actin mRNA. We observed a measurable fluorescence increase in the presence of human total RNA relative to *Escherichia coli* total RNA, demonstrating the specificity of this method (Fig. 4c). Furthermore, given that C2c2 processes its own guide, we combined pre-crRNA processing and RNA detection in a single reaction by designing tandem crRNA-repeat-containing spacers complementary to target RNAs A and  $\lambda$ 2. LbuC2c2 incubated with

this unprocessed tandem guide RNA in the detection assay generated a substantial fluorescence increase similar in magnitude and sensitivity to experiments using mature crRNAs (Fig. 4b, d). Taken together, these data highlight the exciting opportunity to take advantage of the two distinct RNase activities of C2c2 for a range of biotechnological applications (Fig. 4e).

In bacteria, C2c2 probably operates as a sentinel for viral RNAs<sup>9</sup>. We propose that when invasive transcripts are detected within the host cell via base pairing with crRNAs, C2c2 is activated for promiscuous cleavage of RNA in *trans* (Fig. 4e). This bears notable similarity to RNase L and caspase defence mechanisms in eukaryotes, in which a cellular signal triggers promiscuous ribonucleolytic and proteolytic degradation, respectively, within the host cell, leading to apoptosis<sup>28,29</sup>. While the RNA targeting mechanisms of type III CRISPR systems generally result in RNA cleavage within the protospacer–guide duplex<sup>8,30</sup>, recent examples of associated nucleases Csx1 (ref. 23) and Csm6 (ref. 24) provide compelling parallels between the type VI systems and the multi-component type III inference complexes.

Our data show that CRISPR C2c2 proteins represent a class of enzyme capable of two separate RNA recognition and cleavage activities. Efficient pre-crRNA processing requires sequence and structural motifs within the CRISPR repeat that prevent non-endogenous crRNA loading and helps to reduce the potential toxicity of this potent RNase. The entirely different pre-crRNA processing mechanisms of C2c2 and the type V CRISPR effector protein Cpf1 indicate that each protein family has converged upon independent activities encompassing both the processing and interference functions of their respective CRISPR pathways. Furthermore, the two distinct catalytic capabilities of C2c2 can be harnessed together for RNA detection, as the activation of C2c2 to cleave thousands of *trans* RNAs for every target RNA detected enables potent signal amplification. The capacity of C2c2 to process its own guide RNAs from arrays could also allow the use of tissue-specific polymerase II promoters for guide expression, in addition to target multiplexing for a wide range of applications. The C2c2 enzyme is unique within bacterial adaptive immunity for its dual RNase activities, and highlights the utility of harnessing CRISPR proteins for precise nucleic acid manipulation in cells and cell-free systems.



**Figure 4 | C2c2 provides sensitive detection of transcripts in complex mixtures.** **a**, Illustration of LbuC2c2 RNA detection approach using a quenched fluorescent RNA reporter. **b**, Quantification of fluorescence signal generated after 30 min by wild-type or catalytically dead (dHEPN1/2) LbuC2c2 loaded with either a  $\lambda$ 2-,  $\lambda$ 3- or  $\lambda$ 4-targeting crRNA (cr-; as indicated) in the presence of varying concentrations of  $\lambda$ 2– $\lambda$ 4 target ssRNA and human total RNA. RNase A shown as positive RNA degradation control (mean  $\pm$  s.d.,  $n = 3$ ). AU, arbitrary units. **c**, Quantification of fluorescence signal generated by LbuC2c2 loaded with a  $\beta$ -actin targeting crRNA after 3 h for varying amounts of human total RNA or bacterial total RNA (as a  $\beta$ -actin-null negative control) (mean  $\pm$  s.d.,  $n = 3$ ). **d**, Tandem pre-crRNA processing also enables RNA detection (mean  $\pm$  s.d.,  $n = 3$ ). **e**, Model of the type VI CRISPR pathway highlighting both of the C2c2 RNase activities.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 14 June; accepted 24 August 2016.**

**Published online 26 September; corrected online 12 October 2016**

(see full-text HTML version for details).

- van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **12**, 479–492 (2014).
- Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. *Cell* **164**, 29–44 (2016).
- Brouns, S. J. J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960–964 (2008).
- Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843–1845 (2008).
- Garneau, J. E. et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
- Hale, C. R. et al. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**, 945–956 (2009).
- Staals, R. H. J. et al. Structure and activity of the RNA-targeting type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Mol. Cell* **52**, 135–145 (2013).
- Samai, P. et al. Co-transcriptional DNA and RNA cleavage during type III CRISPR-Cas immunity. *Cell* **161**, 1164–1174 (2015).
- Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
- Shmakov, S. et al. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. *Mol. Cell* **60**, 385–397 (2015).

- Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–521 (2016).
- Li, H. Structural principles of CRISPR RNA processing. *Structure* **23**, 13–20 (2015).
- Charpentier, E., Richter, H., van der Oost, J. & White, M. F. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol. Rev.* **39**, 428–441 (2015).
- Hochstrasser, M. L. & Doudna, J. A. Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem. Sci.* **40**, 58–66 (2015).
- Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* **22**, 3489–3496 (2008).
- Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).
- Nam, K. H. et al. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure* **20**, 1574–1584 (2012).
- Deltcheva, E. et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602–607 (2011).
- Fonfara, I. et al. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.* **42**, 2577–2590 (2014).
- Yang, W. Nucleases: diversity of structure, function and mechanism. *Q. Rev. Biophys.* **44**, 1–93 (2011).
- Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
- Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V. & Aravind, L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol. Direct* **8**, 15 (2013).
- Sheppard, N. F., Glover, C. V. C., III, Terns, R. M. & Terns, M. P. The CRISPR-associated Csx1 protein of *Pyrococcus furiosus* is an adenosine-specific endoribonuclease. *RNA* **22**, 216–224 (2016).
- Niewoehner, O. & Jinek, M. Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. *RNA* **22**, 318–329 (2016).
- Cordray, M. S. & Richards-Kortum, R. R. Emerging nucleic acid-based tests for point-of-care detection of malaria. *Am. J. Trop. Med. Hyg.* **87**, 223–230 (2012).
- Rohrman, B. A., Leautaud, V., Molyneux, E. & Richards-Kortum, R. R. A lateral flow assay for quantitative detection of amplified HIV-1 RNA. *PLoS One* **7**, e45611 (2012).
- Yan, L. et al. Isothermal amplified detection of DNA and RNA. *Mol. Biosyst.* **10**, 970–1003 (2014).
- McIlwain, D. R., Berger, T. & Mak, T. W. Caspase functions in cell death and disease. *Cold Spring Harb. Perspect. Biol.* **5**, a008656 (2013).
- Choi, U. Y., Kang, J.-S., Hwang, Y. S. & Kim, Y.-J. Oligoadenylate synthase-like (OASL) proteins: dual functions and associations with diseases. *Exp. Mol. Med.* **47**, e144 (2015).
- Zhang, J., Graham, S., Tello, A., Liu, H. & White, M. F. Multiple nucleic acid cleavage modes in divergent type III CRISPR systems. *Nucleic Acids Res.* **44**, 1789–1799 (2016).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank the QB3 Macrolab for assistance with cloning of C2c2 constructs; N. Ma and K. Zhou for technical assistance; S. N. Floor, S. C. Strutt, A. V. Wright and M. L. Hochstrasser for critical reading of the manuscript; and members of the Doudna, Cate and Tjian laboratories for discussions. S.C.K. acknowledges support from the National Science Foundation Graduate Research Fellowship Program; M.R.O. is a recipient of a C. J. Martin Overseas Early Career Fellowship from the National Health and Medical Research Council of Australia. This work was supported in part by a Frontiers Science award from the Paul Allen Institute to J.A.D., the National Science Foundation (MCB-1244557 to J.A.D.), the California Institute for Regenerative Medicine (CIRM, RB4-06016 to R.T.), and the National Institutes of Health (P50-GM102706 to J.H.D.C.). R.T. and J.A.D. are Investigators of the Howard Hughes Medical Institute. J.A.D. is a co-founder of Caribou Biosciences, Editas Medicine and Intellia Therapeutics and a scientific advisor to Caribou, Intellia, eFFECTOR Therapeutics, and Driver. A.E.S., M.R.O., S.C.K., J.H.D.C. and J.A.D. have filed a patent application related to this work.

**Author Contributions** A.E.S., M.R.O. and S.C.K. conceived the study and designed experiments with input from J.H.D.C., R.T. and J.A.D. D.B. performed bioinformatic analyses. A.E.S. and M.R.O. executed all experimental work with assistance from S.C.K. All authors discussed the data and wrote the manuscript.

**Author Information** Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.A.D. ([doudna@berkeley.edu](mailto:doudna@berkeley.edu)).

**Reviewer Information** Nature thanks M. White, J. Wilusz and the other anonymous reviewer(s) for their contribution to the peer review of this work.

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

**C2c2 phylogenetic and candidate selection.** C2c2 maximum-likelihood phylogenies were computed using RAXML<sup>31</sup> with the PROTGAMMALG evolutionary model and 100 bootstrap samplings. Sequences were aligned by MAFFT with the 'einsi' method<sup>32</sup>.

**C2c2 protein production and purification.** Expression vectors for protein purification were assembled using synthetic gBlocks ordered from Integrated DNA Technologies. The codon-optimized C2c2 genomic sequence was N-terminally tagged with a His<sub>6</sub>-MBP-TEV cleavage site, with expression driven by a T7 promoter. Mutant proteins were cloned via site-directed mutagenesis of wild-type C2c2 constructs. Expression vectors were transformed into Rosetta2 *E. coli* cells grown in 2×YT broth at 37°C. *E. coli* cells were induced during log phase with 0.5 mM IPTG, and the temperature was reduced to 16°C for overnight expression of His-MBP-C2c2. Cells were subsequently harvested, resuspended in lysis buffer (50 mM Tris-HCl pH 7.0, 500 mM NaCl, 5% glycerol, 1 mM TCEP, 0.5 mM PMSF, and EDTA-free protease inhibitor (Roche)) and lysed by sonication, and the lysates were clarified by centrifugation. Soluble His-MBP-C2c2 was isolated over metal ion affinity chromatography, and protein-containing eluate was incubated with TEV protease at 4°C overnight while dialyzing into ion exchange buffer (50 mM Tris-HCl pH 7.0, 250 mM KCl, 5% glycerol, 1 mM TCEP) in order to cleave off the His<sub>6</sub>-MBP tag. Cleaved protein was loaded onto a HiTrap SP column and eluted over a linear KCl (0.25–1.5 M) gradient. Cation exchange chromatography fractions were pooled and concentrated with 30 kDa cutoff concentrators (Thermo Fisher). The C2c2 protein was further purified via size-exclusion chromatography on an S200 column and stored in gel filtration buffer (20 mM Tris-HCl pH 7.0, 200 mM KCl, 5% glycerol, 1 mM TCEP) for subsequent enzymatic assays. Expression plasmids are deposited with Addgene.

**Generation of RNA.** All RNAs used in this study were transcribed *in vitro* except for crRNA AES461, which was ordered synthetically (Integrated DNA Technologies) (see Extended Data Table 1, Extended Data Fig. 6). *In vitro* transcription reactions were performed as previously described with the following modifications: the T7 polymerase concentration was reduced to 10 µg ml<sup>-1</sup>, and the UTP concentration was reduced to 2.5 mM<sup>33</sup>. Transcriptions were incubated at 37°C for 1–2 h to reduce non-template addition of nucleotides and quenched via treatment with DNase I at 37°C for 0.5–1 h. Transcription reactions were purified by 15% denaturing polyacrylamide gel electrophoresis (PAGE), and all RNAs were resuspended in cleavage buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM MgCl<sub>2</sub>, and 5% glycerol). For radioactive experiments, 5' triphosphates were removed by calf intestinal phosphatase (New England Biolabs) before radiolabelling and ssRNA substrates were then 5'-end labelled using T4 polynucleotide kinase (New England Biolabs) and [ $\gamma$ -<sup>32</sup>P]ATP (Perkin Elmer) as described previously<sup>33</sup>.

**Pre-crRNA processing assays.** Pre-crRNA cleavage assays were performed at 37°C in RNA processing buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM MgCl<sub>2</sub>, 10 µg ml<sup>-1</sup> BSA, 10 µg ml<sup>-1</sup> tRNA, 0.05% Igepal CA-630 and 5% glycerol) with a 100-fold molar excess of C2c2 relative to 5'-labelled pre-crRNA (final concentrations of 100 nM and <1 nM, respectively). Unless otherwise indicated, the reaction was quenched after 1 h with 1.5× RNA loading dye (100% formamide, 0.025% (w/v) bromophenol blue and 200 µg ml<sup>-1</sup> heparin). After quenching, reactions were denatured at 95°C for 5 min before resolving by 12% or 15% denaturing PAGE (0.5× TBE buffer). Metal dependence of the reaction was tested by addition of EDTA or EGTA to the reaction buffer at concentrations varying from 10 to 100 mM. Bands were visualized by phosphorimaging and quantified with ImageQuant (GE Healthcare). The percentage cleavage was determined as the ratio of the product band intensity to the total intensity of both the product and uncleaved pre-crRNA bands and normalized for background within each measured substrate using ImageQuant TL Software (GE Healthcare) and fit to a one phase exponential association using Prism (GraphPad).

**Product size mapping and 3' end moiety identification.** Cleavage product length was determined biochemically by comparing gel migration of product bands to alkaline hydrolysis and RNase T1 digestion ladders using the RNase T1 Kit from Ambion. For the hydrolysis ladder, 15 nM full-length RNA substrates were incubated at 95°C in 1× alkaline hydrolysis buffer (Ambion) for 5 min. Reactions were quenched with 1.5× RNA loading buffer, and cooled to -20°C to immediately stop hydrolysis. For the RNase T1 ladder, 15 nM full-length RNA substrates were unfolded in 1× RNA sequencing buffer (Ambion) at 65°C. Reactions were cooled to ambient temperature, and then 1 U of RNase T1 (Ambion) was added to reaction. After 15 min, reactions were stopped by phenol–chloroform extraction and 1.5× RNA loading buffer was added for storage. Hydrolysis bands were resolved in parallel to cleavage samples on 15% denaturing PAGE and visualized by phosphorimaging.

For 3' end moiety identification, products from the processing reaction were incubated with 10 U of T4 polynucleotide kinase (New England Biolabs) for 1 h at 37°C in processing buffer. Reactions were quenched with 1.5× RNA loading buffer, resolved on 20% denaturing PAGE and visualized by phosphorimaging.

**Small RNA sequencing analysis.** RNA reads from ref. 10 were downloaded from SRA runs SRR3713697, SRR3713948 and SRR3713950. The paired-end reads were locally mapped to the reference sequences using Bowtie2 (ref. 34) with the following options: “-reorder-very-fast-local-local”. The mapping was then filtered to retain only alignments that contained no mismatch using mapped.py (<https://github.com/christophertbrown/mapped>) with the “-m 0 -p both” options. BAM files of the resulting mapping are in the Supplementary Information. Read coverage was visualized using Geneious and plotted using Prism (GraphPad).

**Target cleavage assays.** Target cleavage assays were performed at 25°C or 37°C in cleavage buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM MgCl<sub>2</sub> and 5% glycerol). crRNA guides were pre-folded by heating to 65°C for 5 min and then slowly cooling to ambient temperature in cleavage buffer. C2c2–crRNA complex formation was performed in cleavage buffer, generally at a molar ratio of 2:1 protein to crRNA at 37°C for 10 min, before adding 5'-end labelled target and/or other non-radiolabelled RNA target substrates. Unless otherwise indicated, final concentrations of protein, guide and targets were 100 nM, 50 nM and <1 nM, respectively, for all reactions. Reactions were quenched with 1.5× RNA loading dye and resolved by 15% denaturing PAGE (0.5× TBE buffer). Bands were visualized by phosphorimaging and quantified with ImageQuant (GE Healthcare). The percentage cleavage was determined as the ratio of total banding intensity for all shorter products relative to the uncleaved band and normalized for background within each measured substrate using ImageQuant TL Software (GE Healthcare) and fit to a one phase exponential association using Prism (GraphPad).

**crRNA filter-binding assays.** Filter binding assays were carried out in RNA processing buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM MgCl<sub>2</sub>, 10 µg ml<sup>-1</sup> BSA, 10 µg ml<sup>-1</sup> yeast tRNA, 0.01% Igepal CA-630 and 5% glycerol). LbuC2c2 was incubated with radiolabelled crRNA (<0.1 nM) for 1 h at 37°C. Tuffryn, Protran and Hybond-N+ were assembled onto a dot-blot apparatus in the order listed above. The membranes were washed twice with 50 µl equilibration buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM MgCl<sub>2</sub> and 5% glycerol) before the sample was applied to the membranes. Membranes were again washed with 50 µl equilibration buffer, dried and visualized by phosphorimaging. Data were quantified with ImageQuant TL Software (GE Healthcare) and fit to a binding isotherm using Prism (GraphPad Software). All experiments were carried out in triplicate. Dissociation constants and associated errors are reported in the figure legends.

**Electrophoretic mobility-shift assays.** To avoid the dissociation of the LbuC2c2-dHEPN1/dHEPN2–crRNA complex at low concentrations during ssRNA-binding experiments, binding reactions contained a constant excess of LbuC2c2-dHEPN1/dHEPN2 (200 nM), and increasing concentrations of crRNA-A and <0.1 nM target ssRNA. Assays were carried out in C2c2 EMSA buffer (20 mM HEPES pH 6.8, 50 mM KCl, 10 µg ml<sup>-1</sup> BSA, 100 µg ml<sup>-1</sup> yeast tRNA, 0.01% Igepal CA-630 and 5% glycerol). LbuC2c2–crRNA-A complexes were pre-formed as described above for 10 min at 37°C before the addition of 5'-radiolabelled ssRNA substrate and a further incubation for 45 min at 37°C. Samples were then resolved by 8% native PAGE at 4°C (0.5× TBE buffer). Gels were imaged by phosphorimaging, quantified using ImageQuant TL Software (GE Healthcare) and fit to a binding isotherm using Prism (GraphPad Software). All experiments were carried out in triplicate. Dissociation constants and associated errors are reported in the figure legends.

**Fluorescent RNA detection assay.** LbuC2c2–crRNA complexes were preassembled by incubating 1 µM of LbuC2c2 with 500 nM of crRNA for 10 min at 37°C. These complexes were then diluted to 100 nM LbuC2c2, 50 nM crRNA, in RNA processing buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM MgCl<sub>2</sub>, 10 µg ml<sup>-1</sup> BSA, 10 µg ml<sup>-1</sup> yeast tRNA, 0.01% Igepal CA-630 and 5% glycerol) in the presence of 185 nM of RNAase-Alert substrate (Thermo-Fisher), 100 ng of HeLa total RNA and increasing amounts of target 60-nucleotide ssRNA (0–1 nM). These reactions were incubated in a fluorescence plate reader (Tecan Infinite Pro F200) for up to 120 min at 37°C with fluorescence measurements taken every 5 min ( $\lambda_{\text{ex}}$ : 485 nm;  $\lambda_{\text{em}}$ : 535 nm). Background-corrected fluorescence values were obtained by subtracting fluorescence values obtained from reactions carried out in the absence of target ssRNA. Maximal fluorescence was measured by incubating 50 nM RNase A with 185 nM of RNAase-Alert substrate. For measurement of crRNA-ACTB-mediated LbuC2c2 activation by  $\beta$ -actin mRNA in human total RNA, LbuC2c2–crRNA complexes were preassembled by incubating 1 µM of LbuC2c2 with 500 nM of crRNA-ACTB for 10 min at 37°C and reactions were carried out in the conditions above in the presence of increasing amounts (0–1 µg) of either HeLa cell total RNA or *E. coli* total RNA (as a negative control). These reactions were incubated in a fluorescence plate reader for up to 180 min at 37°C with fluorescence measurements taken every 5 min ( $\lambda_{\text{ex}}$ : 485 nm;  $\lambda_{\text{em}}$ : 535 nm). Background-corrected fluorescence



values were obtained by subtracting fluorescence values obtained from reactions carried out in the absence of target ssRNA. For coupled pre-crRNA processing and RNA detection assays, LbuCas9–crRNA complexes were preassembled by incubating 1  $\mu$ M of LbuC2c2 with 500 nM of pre-crRNA-A- $\lambda$ 2 for 20 min at 37°C and reactions carried out as described above in the presence of increasing amounts of ssRNA A and ssRNA  $\lambda$ 2 (0–1 nM each). In each case, error bars represent the standard deviation from three independent experiments.

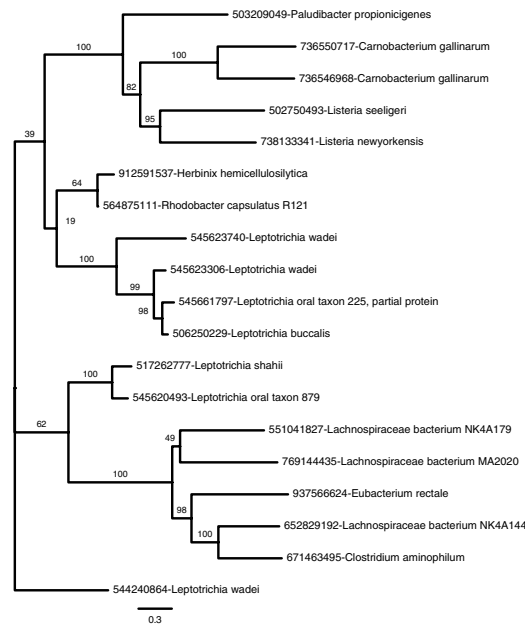
**Background cleavage in total RNA.** LbuC2c2–crRNA- $\lambda$ 4 complexes were assembled as previously described for fluorescence RNA detection assay. Complexes were incubated in RNA processing buffer in the presence of 3  $\mu$ g total RNA with and without 10 nM  $\lambda$ 4 ssRNA target. After 2 h, RNA was isolated by trizol extraction and ethanol precipitation. The RNA fragment size distribution of resuspended samples was resolved using Small RNA Analysis

Kit (Agilent) on a Bioanalyzer 2100 (Agilent) using the manufacturer's protocol. Fluorescent intensity curves were normalized in Prism for curve overlay (GraphPad Software).

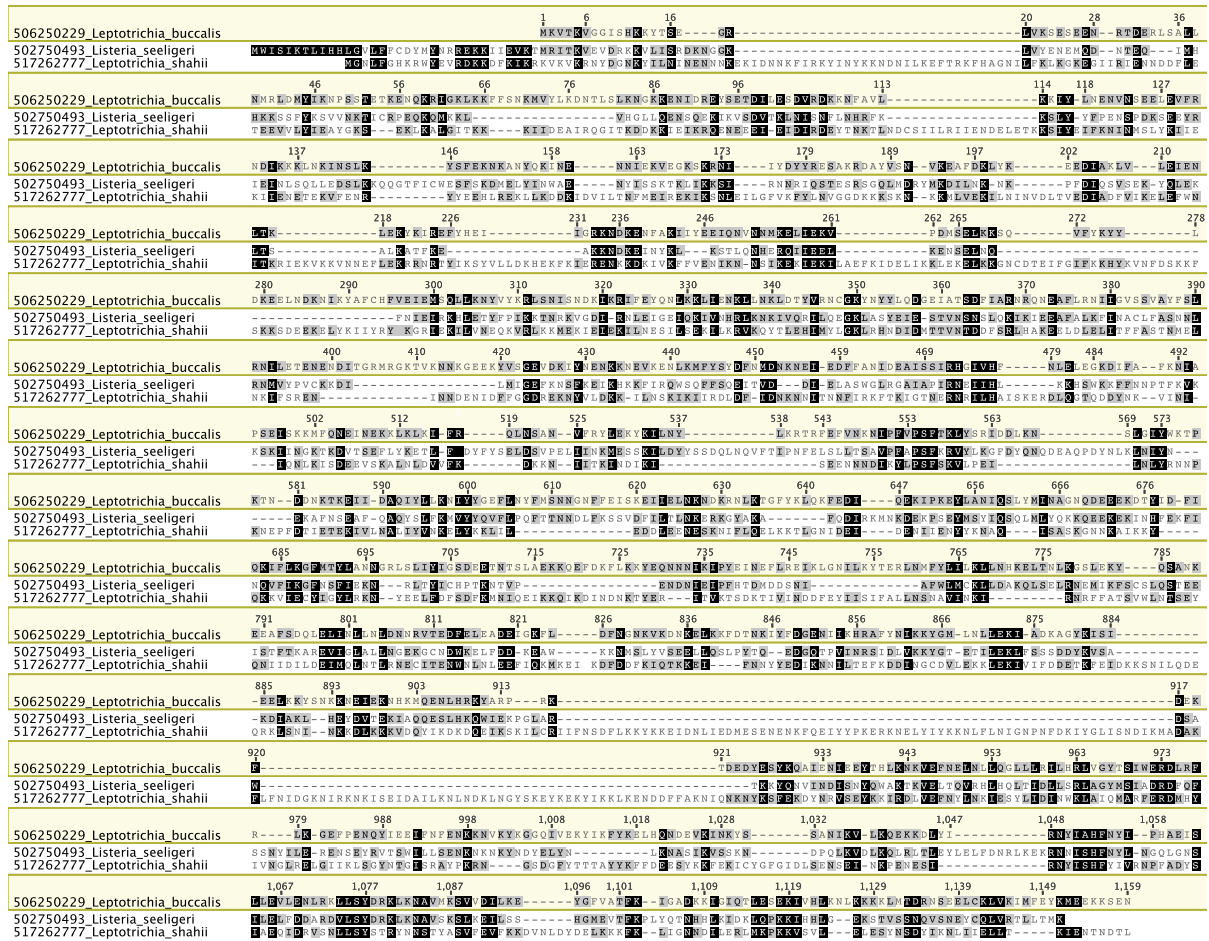
31. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
32. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
33. Sternberg, S. H., Haurwitz, R. E. & Doudna, J. A. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* **18**, 661–672 (2012).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).



a

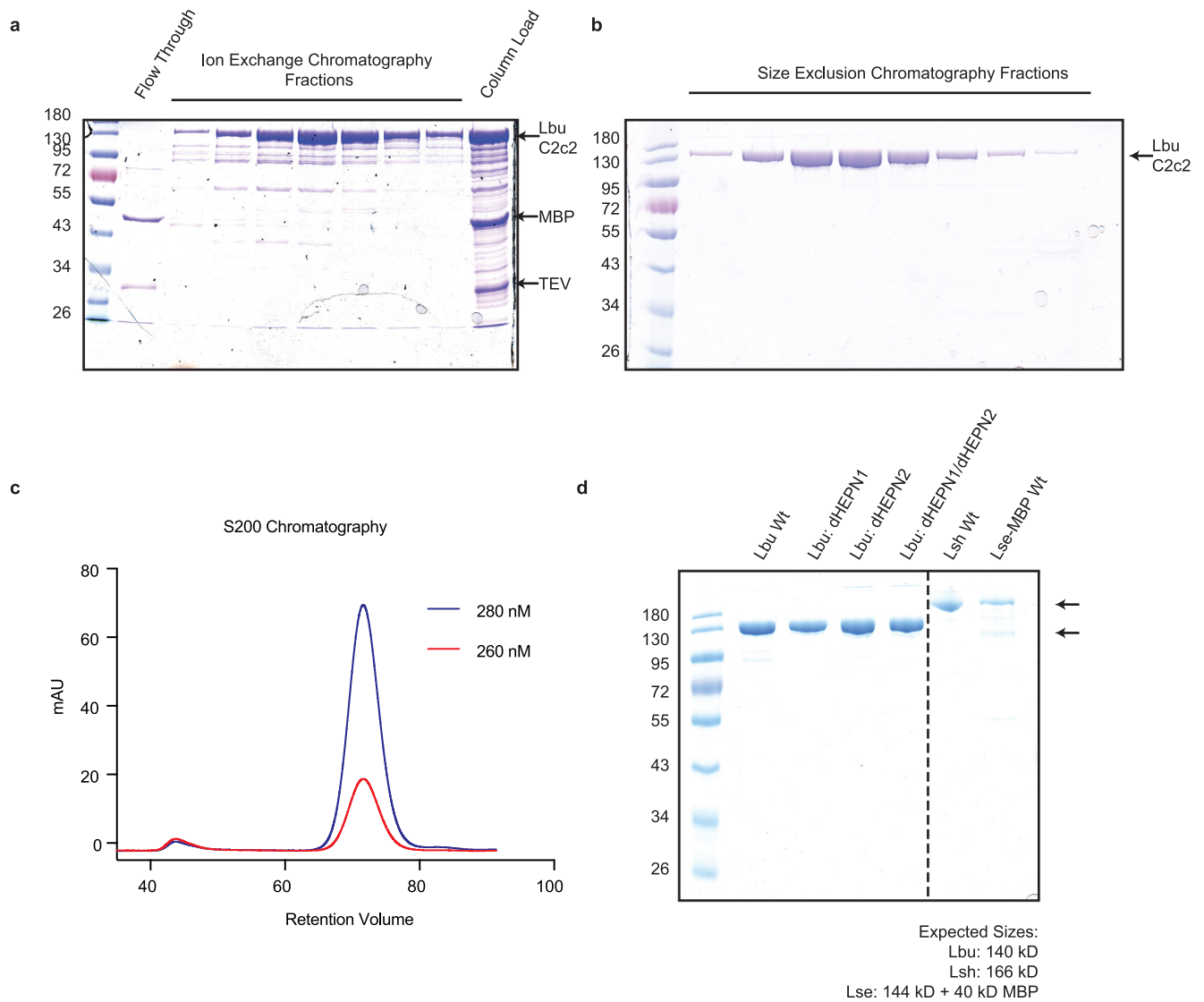


b



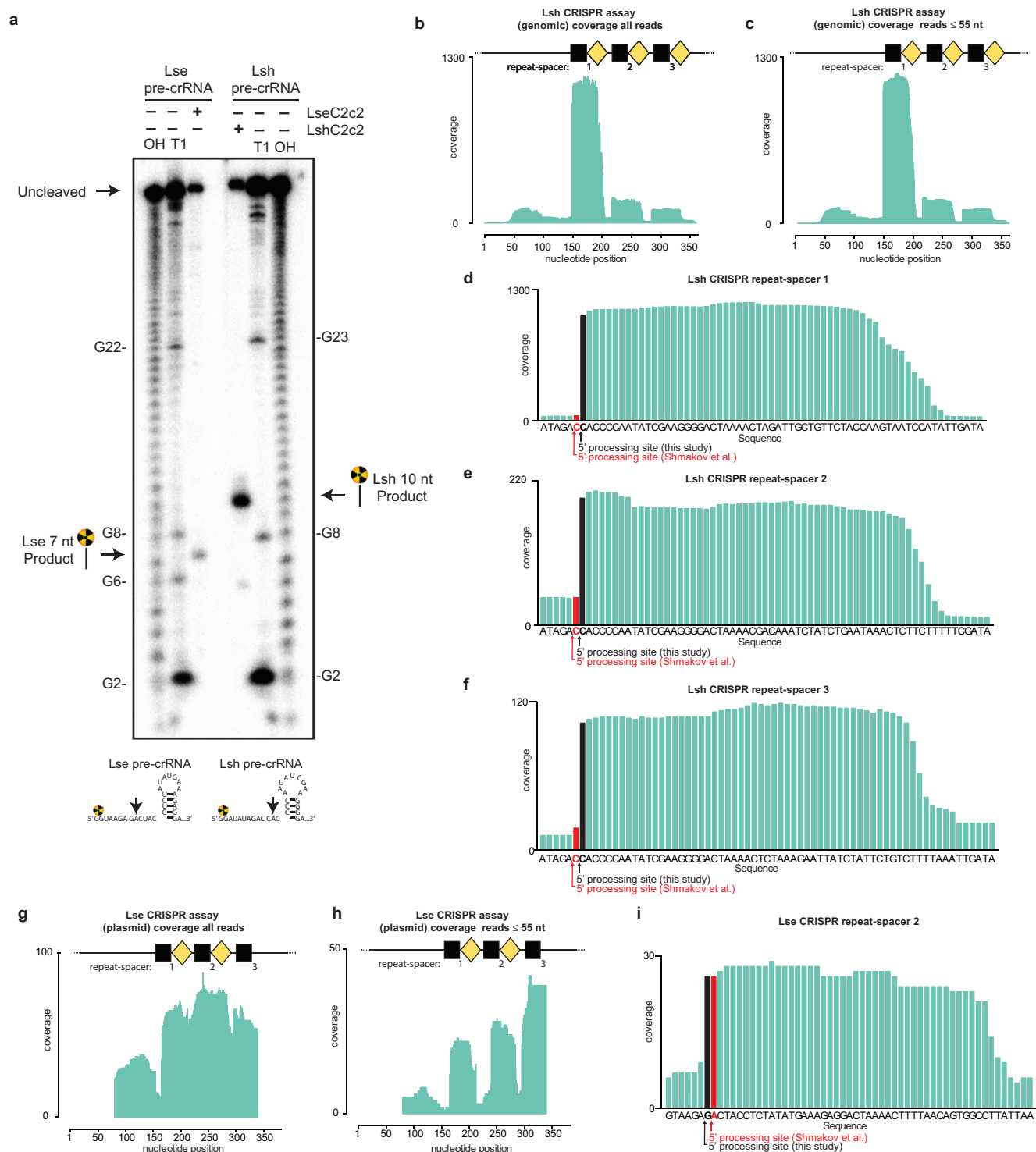
**Extended Data Figure 1 | Complete phylogenetic tree of C2c2 family and C2c2 alignment. a,** Maximum-likelihood phylogenetic reconstruction of C2c2 proteins. Leaves include GI protein numbers and organism of origin; bootstrap support values, out of 100 resamplings, are presented

for inner split. Scale is in substitutions per site. **b,** Multiple sequence alignment of the three analysed homologues of C2c2; coordinates are based on LbuC2c2.



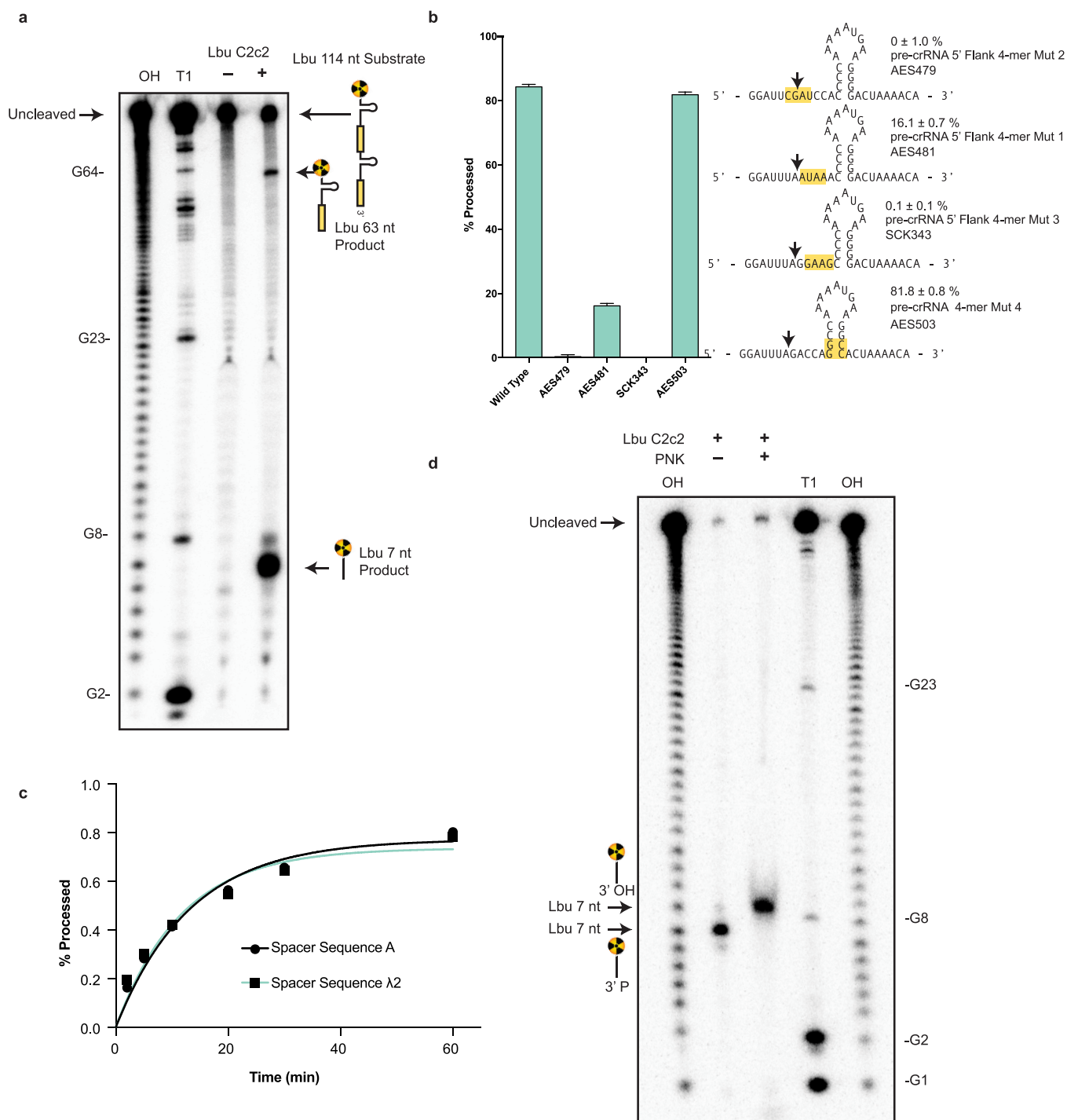
**Extended Data Figure 2 | Purification and Production of C2c2.** All C2c2 homologues were expressed in *E. coli* as His-MBP fusions and purified by a combination of affinity, ion exchange and size exclusion chromatography. The Ni<sup>2+</sup> affinity tag was removed by incubation with TEV protease.

**a, b**, Representative SDS-PAGE gels of chromatography fractions are shown. **c**, The chromatogram from Superdex 200 (16/60) column demonstrating that C2c2 elutes as a single peak, devoid of nucleic acid. **d**, SDS-PAGE analysis of purified proteins used.



**Extended Data Figure 3 | Mapping of pre-crRNA processing by C2c2 *in vitro* and *in vivo*.** **a**, Cleavage site mapping of LseC2c2 and LshC2c2 cleavage of a single cognate pre-crRNA array. Cleavage reactions were performed with 100 nM C2c2 and <1 nM pre-crRNA. **b–i**, Re-analysis of LshC2c2 (**b–f**) and LseC2c2 (**g–i**) CRISPR array RNA sequencing experiments from ref. 10 (supplementary figs S7 and S5 of ref. 10, respectively). All reads (**b, g**) and filtered reads (55 nucleotides or less; as per original analysis<sup>10</sup>; **c, h**) were stringently aligned to each CRISPR array using Bowtie2 (see Methods). Detailed views of individual CRISPR repeat-spacers are shown for Lsh (**d–f**) and Lse (**i**). Differences in 5' end pre-crRNA processing are indicated by arrows below each sequence. BAM alignment files of our analysis are available in Supplementary Information.

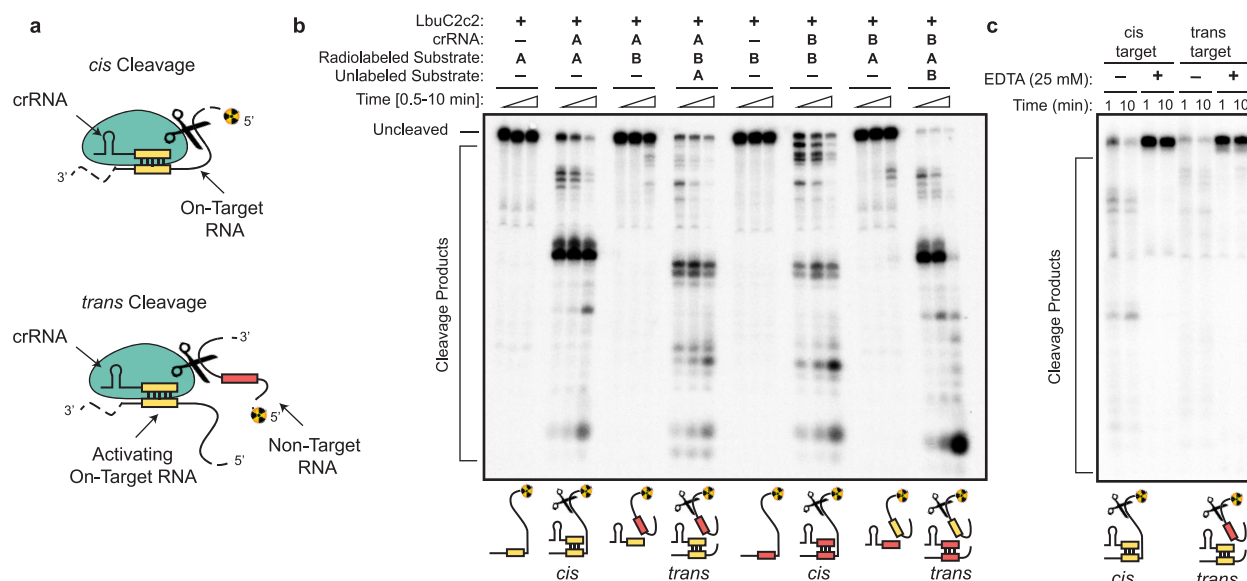
This mapping clearly indicates that the 5' ends of small RNA sequencing reads generated from Lsh pre-crRNAs map to a position 2 nucleotides from the base of the predicted hairpin, in agreement with our *in vitro* processing data (**a**). This pattern holds for all mature crRNAs detected from both native expression in *L. shahii* and heterologous expression in *E. coli* (data not shown, BAM file available in Supplementary Information). Unfortunately, the LseC2c2 crRNA sequencing data (used in **g–i**) is less informative owing to low read depth, and each aligned crRNA exhibits a slightly different 5' end with little obvious uniformity. The mapping for one of the processed repeats (repeat-spacer 2; **i**) is in agreement with our data but only with low confidence due to the insufficient read depth.



**Extended Data Figure 4 | Pre-crRNA processing by C2c2 is spacer-sequence independent, can occur on tandem crRNA arrays, is affected by mutations in the 5' flanking region of the pre-crRNA and produces a 3' phosphate product.** **a**, Cleavage site mapping of LbuC2c2 cleavage of a tandem pre-crRNA array. Cleavage reactions were performed with 100 nM LbuC2c2 and <1 nM pre-crRNA. A schematic of cleavage products is depicted on right, with arrows indicating the mapped C2c2 cleavage products. **b**, LbuC2c2 4-mer mutant pre-crRNA processing data demonstrating the importance of the 5' single-stranded flanking region for efficient pre-crRNA processing. Percentage of pre-crRNA processing was measured after 1 h (mean ± s.d.,  $n = 3$ ). **c**, Representative LbuC2c2

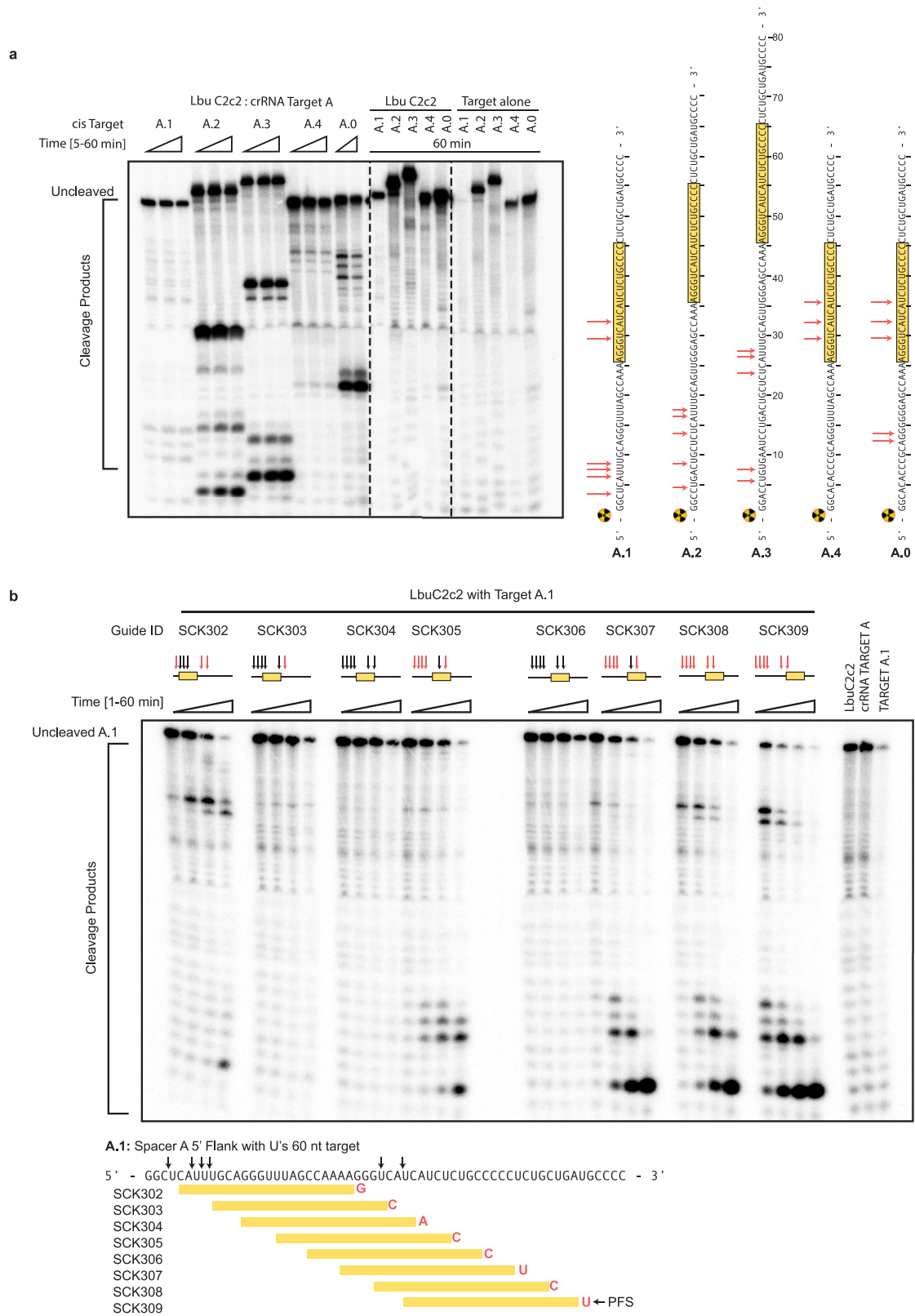
pre-crRNA cleavage time course demonstrating that similar rates of pre-crRNA processing occur independent of crRNA spacer sequence pseudo-first-order rate constants ( $k_{obs}$ ) (mean ± s.d.) are  $0.07 \pm 0.04 \text{ min}^{-1}$  and  $0.08 \pm 0.04 \text{ min}^{-1}$  for spacer A and spacer  $\lambda_2$ , respectively. **d**, End group analysis of cleaved RNA by T4 polynucleotide kinase (PNK) treatment. Standard processing assay conditions were used to generate cleavage product, which was then incubated with PNK for 1 h to remove any 2',3'-cyclic phosphates/3' monophosphates. Retarded migration of band indicates removal of the charged, monophosphate from the 3' end of radiolabelled 5' product.





**Extended Data Figure 5 | LbuC2c2 catalyses guide-dependent ssRNA degradation on *cis* and *trans* targets.** **a**, Schematic of the two modes of C2c2, guide-dependent ssRNA degradation. **b**, Cleavage of two distinct radiolabelled ssRNA substrates, A and B, by LbuC2c2. Complexes of 100 nM C2c2 and 50 nM crRNA were pre-formed at 37 °C, and reaction was initiated upon addition of <1 nM 5'-labelled target RNA at 25 °C. *Trans* cleavage reactions contained equimolar (<1 nM) concentrations of radiolabelled non-guide-complementary substrate, and unlabelled on-target ssRNA. For multiple ssRNA substrates, we observed that LbuC2c2 catalysed efficient cleavage only when bound to the complementary crRNA, indicating that LbuC2c2–crRNA cleaves ssRNA in an RNA-guided fashion. This activity is hereafter referred to as on-target or

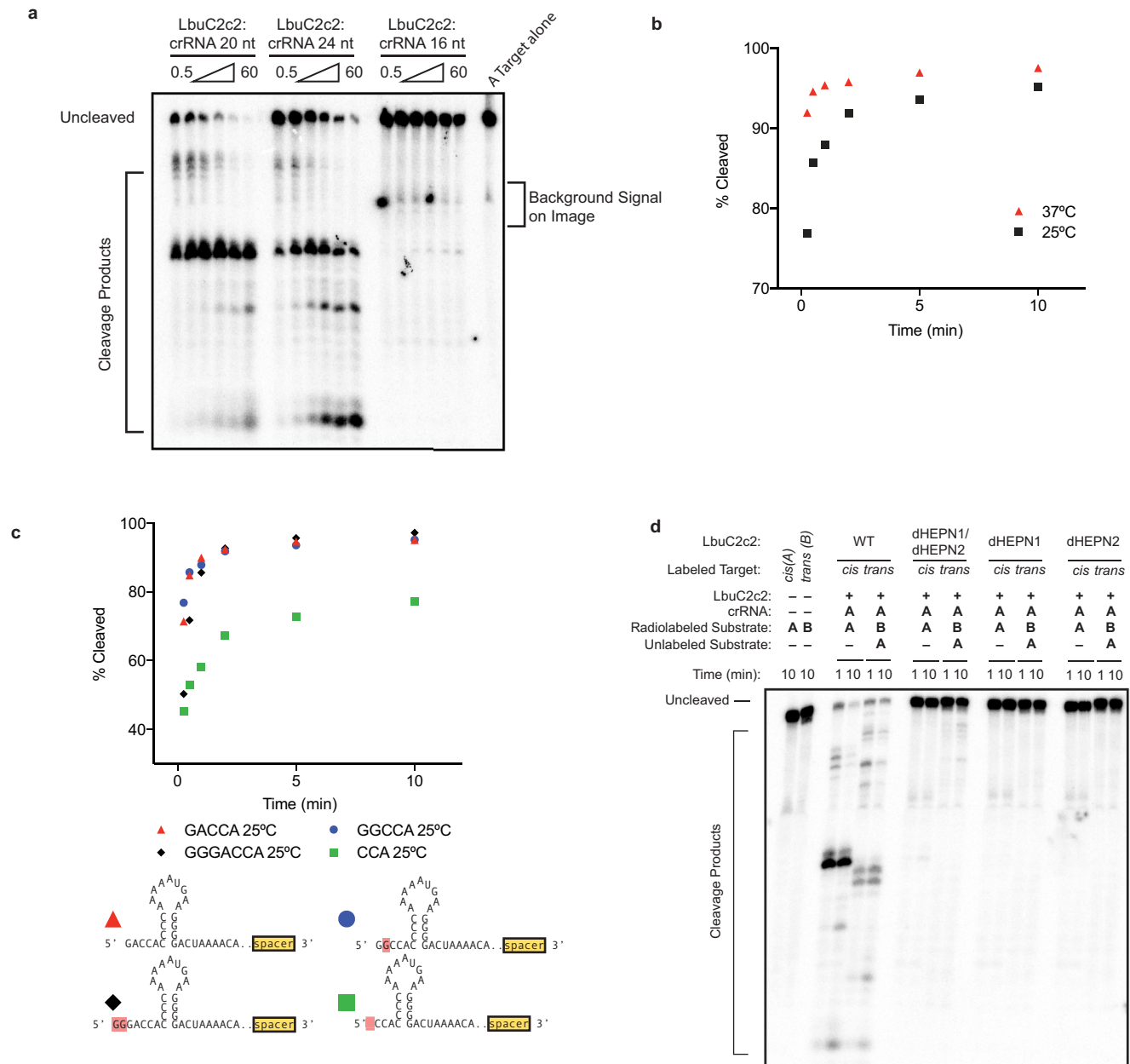
*cis*-target cleavage. LbuC2c2-mediated *cis* cleavage resulted in a laddering of multiple products, with cleavage preferentially occurring before uracil residues, analogous to LshC2c2 (ref. 9). We repeated non-target cleavage reactions in the presence of unlabelled, on-target (crRNA-complementary) ssRNA. In contrast to non-target cleavage experiments performed in *cis*, we observed rapid degradation of non-target RNA in *trans*. The similar RNA cleavage rates and near-identical cleavage products observed for both *cis* on-target cleavage and *trans* non-target cleavage implicate the same nuclease centre in both activities. **c**, LbuC2c2 loaded with crRNA targeting spacer A was tested for cleavage activity under both *cis* (target A labelled) and *trans* (target B labelled in the presence of unlabelled target A) cleavage conditions in the presence of 25 mM EDTA.



Extended Data Figure 6 | See next page for caption.

**Extended Data Figure 6 | LbuC2c2 ssRNA target cleavage site mapping.** **a**, ssRNA target cleavage assay conducted per Methods demonstrating LbuC2c2-mediated ‘*cis*’ cleavage of several radiolabelled ssRNA substrates with identical spacer-complementary sequences but distinct 5′ flanking sequences of variable length and nucleotide composition. Sequences of ssRNA substrates are shown to the right with spacer-complementary sequences for crRNA-A highlighted in yellow. Arrows indicate detected cleavage sites. Gel was cropped for clarity. It should be noted that the pattern of cleavage products produced on different substrates (for example, A.1 versus A.2 versus A.3) indicates that the cleavage site choice is primarily driven by a uracil preference and exhibits an apparent lack of exclusive cleavage mechanism within the crRNA-complementary target sequence, which is in contrast to what is observed for other class II CRISPR single effector complexes such as Cas9 and Cpf1 (refs 11, 21). Notably, the cleavage pattern observed for substrate A.0 hints at a secondary preference for polyG sequences. **b**, LbuC2c2 ssRNA target cleavage assay as per Methods, using a range of crRNAs that tile the length

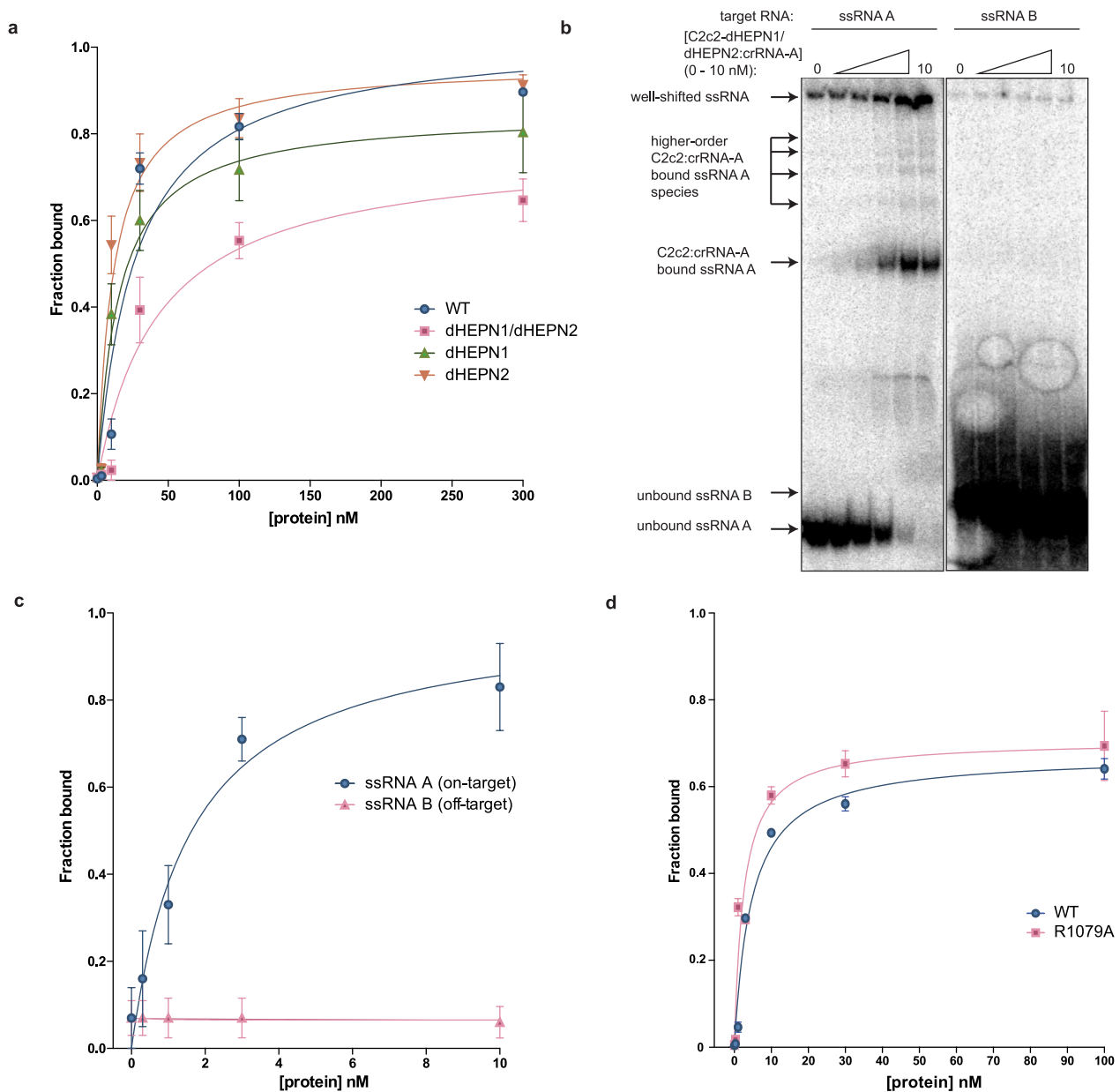
of the ssRNA target. The sequence of the ssRNA substrates used in this experiment is shown below the gel with spacer-complementary sequences for each crRNA highlighted in yellow. Arrows indicate predicted cleavage sites. Above each set of lanes, a small diagram indicates the location of the spacer sequence along the target (yellow box) and the cleavage products observed (red arrows) or absent (black arrows). Likewise, it should be noted that for every crRNA the cleavage product length distribution is very similar, again indicating an apparent lack of exclusive cleavage within the crRNA-bound sequence. The absence of a several cleavage products in a subset of the reactions might be explained by the presence of bound C2c2–crRNA on the ssRNA target, which could sterically occlude access to uracils by any *cis* (intramolecular) or *trans* (intermolecular) LbuC2c2 active sites. While proper analysis for protospacer flanking site (PFS) preference for LbuC2c2 is beyond the scope of this study, minimal impact of the 3′ flanking nucleotide was observed. Expected PFS base is noted in diagram next to each guide tested in red.



**Extended Data Figure 7 | Dependence of RNA targeting on crRNA variants, temperature and point mutations.** **a**, LbuC2c2 ssRNA target cleavage assay carried out, as per Methods with crRNAs possessing 16-, 20- or 24-nucleotide spacers. **b**, LbuC2c2 ssRNA target cleavage time-course carried out at either 25°C or 37°C as per the Methods. **c**, LbuC2c2 ssRNA target cleavage time course carried out as stated in the Methods

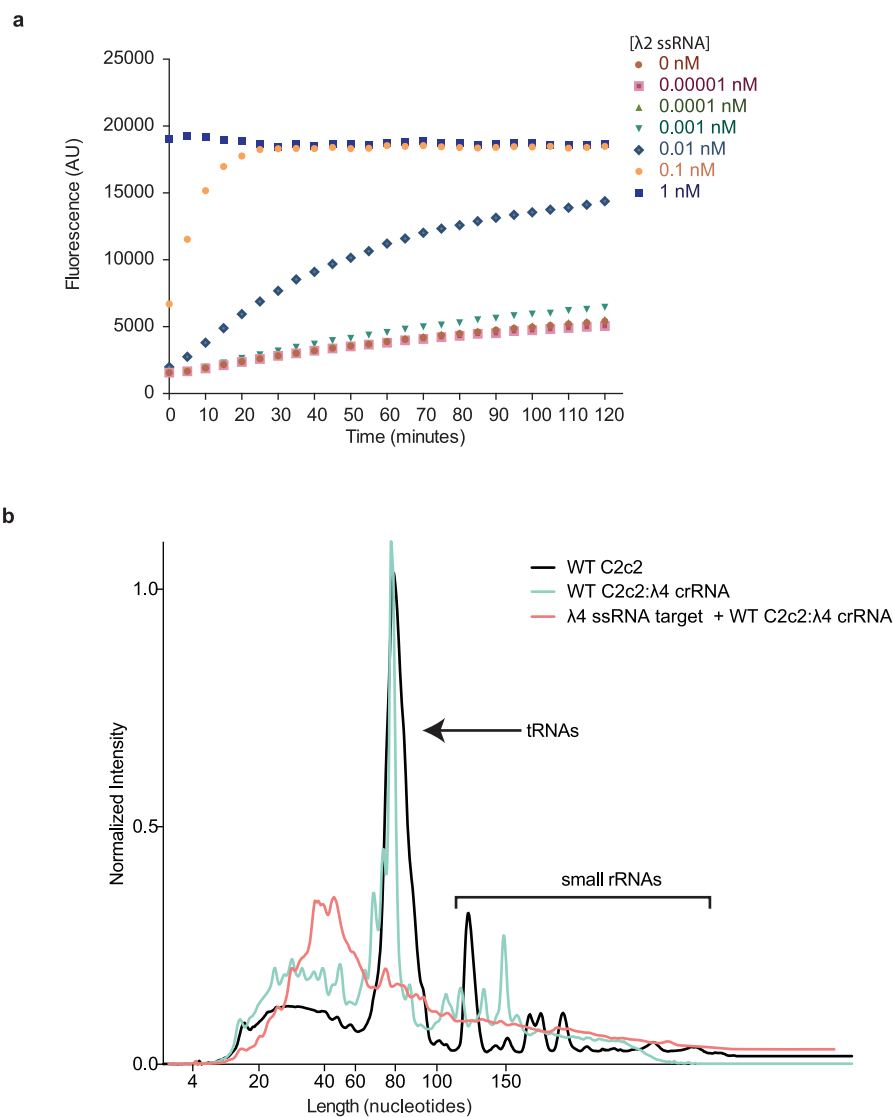
with crRNAs possessing different 5'-flanking nucleotide mutations. Mutations are highlighted in red. One- to two-nucleotide 5' extensions negligibly impacted cleavage efficiencies. By contrast, shortening the flanking region to 3 nucleotides slowed cleavage rates. **d**, Effect of point mutations on RNase activity of C2c2 in conserved residue mutants within HEPN motifs for ssRNA targeting.





**Extended Data Figure 8 | Binding data for LbuC2c2 to mature crRNA and target ssRNA.** **a**, Filter binding assays were conducted as described in the Methods to determine the binding affinity of mature crRNA-A<sub>GG</sub> to LbuC2c2-WT, LbuC2c2-dHEPN1, LbuC2c2-dHEPN2, or LbuC2c2-dHEPN1/dHEPN2. The quantified data were fit to standard binding isotherms. Error bars represent the s.d. from three independent experiments. Measured dissociation constants from three independent experiments (mean  $\pm$  s.d.) were  $27.1 \pm 7.5$  nM (LbuC2c2-WT),  $15.2 \pm 3.2$  nM (LbuC2c2-dHEPN1),  $11.5 \pm 2.5$  nM (LbuC2c2-dHEPN2), and  $43.3 \pm 11.5$  nM (LbuC2c2-dHEPN1/dHEPN2). **b**, Representative electrophoretic mobility shift assay for binding reactions between LbuC2c2-dHEPN1/dHEPN2: crRNA-A<sub>GG</sub> and either 'on-target' A ssRNA or 'off-target' B ssRNA, as indicated. Three independent experiments were conducted as described in the Methods. The gel was cropped for clarity. **c**, Quantified binding data from **b** were fitted to standard binding isoforms. Error bars represent the s.d. from three

independent experiments. Measured dissociation constants from three independent experiments (mean  $\pm$  s.d.) were  $1.62 \pm 0.43$  nM for ssRNA A and not determined (N.D.;  $> 10$  nM) for ssRNA B. **d**, Filter binding assays were conducted as described in the Methods to determine the binding affinity of mature crRNA-A<sub>GA</sub> to LbuC2c2-WT and LbuC2c2-R1079A. The quantified data were fit to standard binding isotherms. Error bars represent the s.d. from three independent experiments. Measured dissociation constants from three independent experiments (mean  $\pm$  s.d.) were  $4.65 \pm 0.6$  nM (LbuC2c2-WT) and  $2.52 \pm 0.5$  nM (LbuC2c2-R1079A). It is of note that these binding affinities differ from **a**. This difference is accounted for in a slight difference in the 5' sequence of the guide with panel **a** guides beginning with a 5'-G(G)CCA... and panel **d** 5'-G(A)CCA. While the native sequence guide (5'-G(A)CCA) binds tighter to LbuC2c2, no difference is seen in the RNA targeting efficiencies of these guide variants (Extended Data Fig. 6c).



**Extended Data Figure 9 | RNase detection assay  $\lambda 2$ -ssRNA time course.** **a**, LbuC2c2:crRNA- $\lambda 2$  was incubated with RNAase-Alert substrate (Thermo-Fisher) and 100 ng HeLa total RNA in the presence of increasing amounts of  $\lambda 2$  ssRNA (0–1 nM) for 2 h at 37 °C. Fluorescence measurements were taken every 5 min. The 1 nM  $\lambda 2$  ssRNA reaction reached saturation before the first time point could be measured. Error bars represent the s.d. from three independent experiments. **b**, LbuC2c2–crRNA- $\lambda 4$  or apo LbuC2c2 was incubated in HeLa total RNA for 2 h in

the presence or absence of on-target activating  $\lambda 4$  ssRNA. Degradation of background small RNA was resolved on a small RNA chip in a Bioanalyzer 2100 as described in the Methods. Small differences are seen in the fragment profile of between apo LbuC2c2 and LbuC2c2:crRNA- $\lambda 4$ . By contrast, upon addition of the on-target ssRNA to the reaction, a notable broadening and shifting of the tRNA peak reveals extensive degradation of other structured and nonstructured RNAs present in the reaction upon activation of LbuC2c2 *trans* activity.

Extended Data Table 1 | Oligonucleotides used in this study

Name	Sequence	Used in
Lbu_pre-crRNA_A_SCK314	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 1c, 2a-c, 3a,c, ED3
Lse_pre-crRNA_B_AES484	5' -GGUAAGAGACUACCCUUAUUAUGAAGAGGAGCUAAAAACCAACAUGAUCUGGGUUAUC-3'	Fig. 1c, ED3
Lsh_pre-crRNA_A_SCK339	5' -GGAUUUAGACCACCCCAAUUAUCGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 1c, ED3
Lbu_pre-crRNA_invert_SCK321	5' -GGAUUUAGACCAGGGGAAGUAAAAACCCACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2a
Lbu_pre-crRNA_5stem_SCK331	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2a
Lbu_pre-crRNA_7bubble_SCK334	5' -GGAUUUAGACCACCCCAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2a
Lbu_pre-crRNA_5bubble_SCK335	5' -GGAUUUAGACCACCCCAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2a
Lbu_pre-crRNA_3stem_SCK342	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2a
Lbu_pre-cr_5'_mut1_AES497	5' -GGGCUUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre-cr_5'_mut2_AES496	5' -GGAGCUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre-cr_5'_mut3_AES495	5' -GGAUUCAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre-cr_5'_mut4_AES477	5' -GGAUUCGGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre-cr_5'_mut5_AES482	5' -GGAUUUACCCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre-cr_5'_mut6_AES478	5' -GGAUUUUAUCCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre-cr_5'_mut7_AES480	5' -GGAUUUUAGAAACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_cr_5'_mut8_AES498	5' -GGAUUUAGACCGCCCCAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_stem_mut1_AES502	5' -GGAUUUAGACACGCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_stem_mut2_AES501	5' -GGAUUUAGACCACCGCAAAAAUGAAGCGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_stem_mut3_AES500	5' -GGAUUUAGACCACACCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_stem_mut4_AES499	5' -GGAUUUAGACCACCAACAAAAUGAAGUGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_stem_mut5_AES504	5' -GGAUUUAGACCACUCCAAAAUGAAGGAGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_cr_3'_mut1_AES505	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_cr_3'_mut2_AES506	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACGCAAAACAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_cr_3'_mut3_AES507	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAGCAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
Lbu_pre_cr_3'_mut4_AES508	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAGUAGGGGCAGAGAUGAUGACCCU-3'	Fig. 2b, ED4
crLbu_A_GG_AES432	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED4-7
crLbu_B_AES451	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAAACAUGAUCUGGGUUAUC-3'	Fig. 3, ED 4
A0_target_AES450	5' -GGCACACCCGACGGGGGAGCCAAAGGGGCAUCAUCUCUGCCCCACAGCAGAAGCCCC-3'	Fig. 3, ED4-7
B_target_AES452	5' -GGGAACCCCAAGGCCAACCCGAGAGAUGACCCAGAUAUUGUAGACCUUACAACCCCC-3'	Fig. 3, ED4,6
crLbu_Lambda2_AES453	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGUGUAUAGUGGAAUGCCAUG-3'	Fig. 4b, ED 8
crLbu_Lambda3_MOC410	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAUUGGUGACAUUCCGAUAGUG-3'	Fig. 4b
crLbu_Lambda4_MOC411	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGAUUAUGCCUGGUGGUUC-3'	Fig. 4b, ED 8
Lambda2_target_MOC28	5' -GGCUCAAUUUUGACAGCGGUCUAGGCAUUCACUUAUCACUGGCAUCCUCCACUC-3'	Fig. 4b, ED 8
Lambda3_target_MOC36	5' -GGAAAUCAUUAACACCCCGCACUAUCGGAAGUUCACAGCCAGCCGACGACGUU-3'	Fig. 4b
Lambda4_target_MOC37	5' -GGCAUAAAAAUGCGCCGCCUGAACCCACAGGCUUAUUCUGCCACUCUUGUUGUGA-3'	Fig. 4b ED 8
crLbu_betaActin_1_AES451	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGUAGUUGGUGUAUC-3'	Fig. 4c
pre-crLbu_dimer_SCK324	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCUUAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGUGUAUAGUGGGAUUGCCAUG-3'	Fig. 4d, ED3
crLbu_lambda2_SCK315	5' -GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAAACAGUUAAGUGGAAUGCCAUG-3'	ED3
Lbu_pre_cr_5'_4mer1_AES481	5' -GGAUUUUAUAAACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED3
Lbu_pre_cr_5'_4mer2_AES479	5' -GGAUUCGAUCCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED3
Lbu_pre_cr_5'_4mer3_SCK343	5' -GGAUUUAGGAAGCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED3
Lbu_pre_cr_5'_4mer4_AES503	5' -GGAUUUAGACCAGGCCCAAAAAUGAAGGCCACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED3
crLbu_GuideWalk1_SCK302	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACUUAUUGGCUCCCCUGCAAAUGAG-3'	ED5
crLbu_GuideWalk2_SCK303	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACACCCUUUUGGCUCCCCUGCAAAA-3'	ED5
crLbu_GuideWalk3_SCK304	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGAUGACCCUUUUGGCUCCCCUG-3'	ED5
crLbu_GuideWalk4_SCK305	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGAUGAUGACCCUUUUGGCUCCCC-3'	ED5
crLbu_GuideWalk5_SCK306	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGCAGAGAUGAUGACCCUUUUGGCU-3'	ED5
crLbu_GuideWalk6_SCK307	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGGCAGAGAUGAUGACCCUUUUU-3'	ED5
crLbu_GuideWalk7_SCK308	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGAGGGGGCAGAGAUGAUGACCC-3'	ED5
crLbu_GuideWalk8_SCK309	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACUACGACAGAGGGGGCAGAGAUGAUG-3'	ED5
A.1_target_U_MOC279	5' -GGCUCAUUAUGCAGGGGGGAGCCAAAGGGGCAUCAUCUCUGCCCCUCUGCUGAUGCCCC-3'	ED5
A.2_target_70nt_AES447	5' -GGCCUGACUGCUCUCAUUUGCAGUUGGGAGCCAAAAGGGUUAUCAUCUCUGCCCCUCUGCUGAUGCCCC-3'	ED5
A.3_target_80nt_AES448	5' -GGACUCUGUAUCCUGACUGCUCUCAUUUGCAGUUGGGAGCCAAAAGGGUUAUCAUCUCUGCCCCUCUGCUGAUGCCCC-3'	ED5
A.4_5'_ts_shift_AES449	5' -GGCACACCCGAGGGUUGGCCAAAGGGGCAUCAUCUCUGCCCCUCUGCUGAUGCCCC-3'	ED5
crLbu_A_16nt_trunc_SCK282	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGAGAUGAUGACCCU-3'	ED6
crLbu_A_24nt_ext_SCK283	5' -GGCCACCCCAAAAAUGAAGGGGACUAAAAACAGAGGGGGCAGAGAUGAUGACCCU-3'	ED6
crLbu_A_mature_GA_SCK340	5' -GACCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED6
crLbu_A_mature_GGGA_SCK341	5' -GGGACACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED6
crLbu_A_mature_CCA_AES461	5' -CCACCCCAAAAAUGAAGGGGACUAAAAACAGGGGCAGAGAUGAUGACCCU-3'	ED6
T7 Forward (DNA)	5' -TAATACGACTCACTATAGG-3'	N/A

## CORRIGENDUM

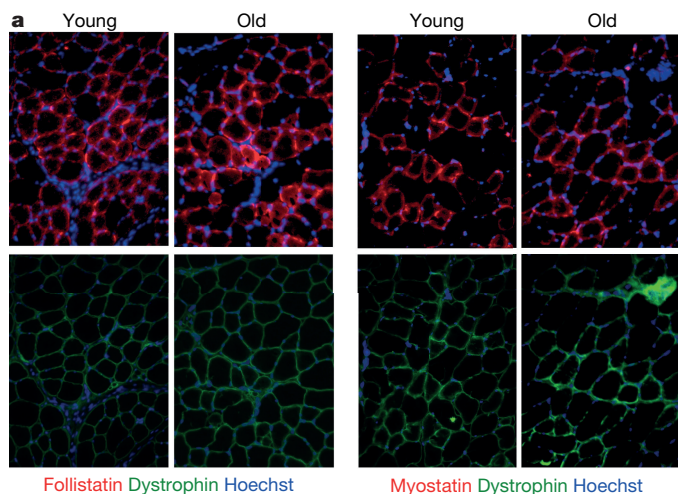
doi:10.1038/nature19077

### Corrigendum: Imbalance between pSmad3 and Notch induces CDK inhibitors in old muscle stem cells

Morgan E. Carlson, Michael Hsu & Irina M. Conboy

*Nature* **454**, 528–532 (2008); doi:10.1038/nature07034

In Fig. 1a of this Letter, the immunofluorescence images for myostatin and follistatin are inaccurate owing to many versions of this figure in multiple revisions of our manuscript. Specifically, the immunofluorescence panels representing anti-myostatin and anti-follistatin staining were duplicated. We would like to thank the anonymous reader who pointed out this error. We have repeated the age-specific immunofluorescence experiments for myostatin and follistatin, and the results confirm our original conclusion that no age-specific differences are detected (Fig. 1). Furthermore, we have performed quantification of the pixel density, and although there are areas with higher and lower signals for these proteins in both young and old muscle sections, the *P* values suggest no age-specific differences. We thank M. Mehdipour for performing the experiments and analysing the data on age-specific muscle tissue levels of myostatin and follistatin.



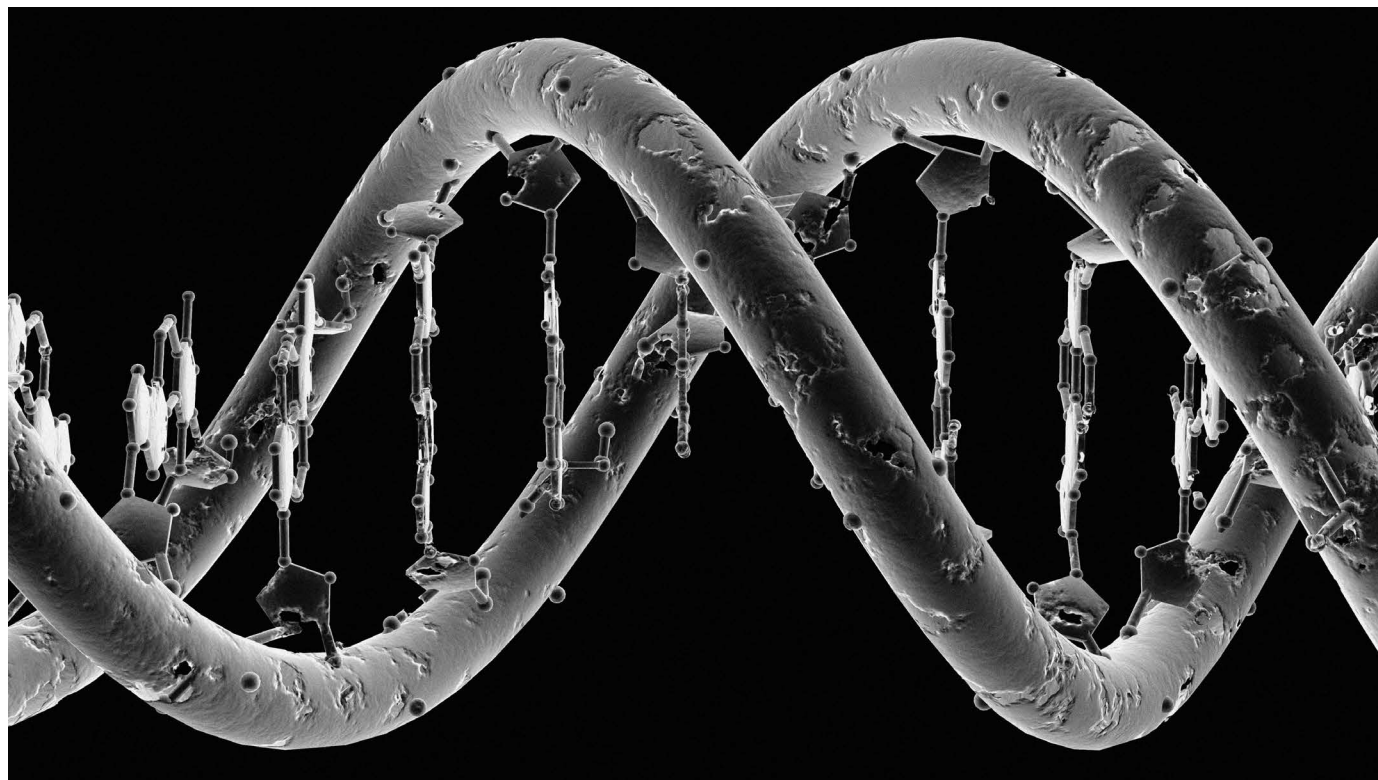
**Figure 1 | This is a repeat of Fig. 1a of the original Letter.** Shown are representative images at  $\times 20$ . Dystrophin immunofluorescence (green) outlines the muscle fibres in 10- $\mu$ m skeletal muscle cryosections that are immunostained for follistatin and myostatin (red); Hoechst (blue) stains all nuclei. *P* values for mean pixel density of myostatin and follistatin (with rabbit IgG control signals subtracted) show no statistically significant differences between young and old muscle.



# THE DARK SIDE OF THE HUMAN GENOME

*Scientists are uncovering the hidden switches in our genome that dial gene expression up and down, but much work lies ahead to peel back the many layers of regulation.*

MEHAU KULYK/SPL



The human genome is not packed with 'junk' as previously thought, but with regulatory regions that modulate gene activity.

BY KELLY RAE CHI

Fifteen years ago, scientists celebrated the first draft of the sequenced human genome. At the time, they predicted that humans had between 25,000 and 40,000 genes that code for proteins. That estimate has continued to fall. Humans actually seem to have as few as 19,000 such genes<sup>1</sup> — a mere 1–2% of the genome. The key to our complexity lies in how these genes are regulated by the remaining 99% of our DNA, known as the genome's 'dark matter'.

From efforts such as the massive Encyclopedia of DNA Elements (ENCODE) project<sup>2</sup>, launched in 2003 by the US National Human Genome Research Institute, it's clear that copious regulatory elements are at play, tuning gene

expression in ways that scientists are only starting to unravel. By uncovering regulatory instructions in the genome beyond protein-coding genes, scientists are hoping to yield new ways to understand and treat disease. "It's not overstating to say that ENCODE is as significant for our understanding of the human genome as the original DNA sequencing of the human genome," says cell biologist Bing Ren of the University of California, San Diego, Institute for Genomic Medicine in La Jolla, who is a member of the ENCODE team.

Ren is also part of a subsequent consortium called the Roadmap Epigenomics Project<sup>3</sup>. These two initiatives — both funded by the US National Institutes of Health (NIH) — aim to map and predict the existence of elements in the genome, including in the vast stretches

of non-coding portions, that drive when and where genes are expressed. Scientists have generated a list of such elements by using biochemical assays to probe DNA sequences, RNA transcripts, regulatory proteins bound to DNA and RNA and epigenetic signatures — the chemical tags on DNA and the proteins packaging it — that also affect gene expression.

So far, the data suggest that there are hundreds of thousands of functional regions in the human genome whose task is to control gene expression: it turns out that much more space in the human genome is devoted to regulating genes than to the genes themselves. Scientists are now trying to validate each predicted element experimentally to ascertain its function — a mammoth task, but one for which they now have a powerful new tool. ►

► Since the gene-editing technique CRISPR–Cas9 entered the scientific arena, the speed at which researchers can test functional elements in the non-coding regions has ramped up. But it is still a daunting endeavour: more than 3 million regulatory DNA regions, thought to contain some 15 million binding sites for regulatory proteins called transcription factors, control gene expression in the human cell types studied thus far. About 150,000 may be active in any given cell type.

These could be crucial to understanding disease, because most single-nucleotide changes associated with common diseases fall in regions outside protein-coding genes, and they often overlap with DNA sites highlighted by ENCODE as having regulatory function. Certain regulatory elements that normally drive gene expression are thought to underpin the mechanism of cancer, for example. Disrupting a gene's regulatory elements, the data suggest, could thus have as drastic an impact on cell function as disrupting the gene itself. Using CRISPR–Cas9, scientists now have an opportunity to test that premise by introducing targeted mutations into non-coding sequences and observing the consequences.

### DECODING A COMPLEX WORLD

How much of DNA's dark matter has a function in gene control is still up for debate. In 2012, ENCODE scientists proposed on the basis of biochemical-assay predictions that 80% of the non-coding genome has a function<sup>2</sup>. But this figure soon proved to be an overestimate as researchers narrowed the definition of 'function' and devised experimental methods, such as reporter assays, to test these functions. "The number still isn't fully known", in part because the mapping isn't complete, says Michael Snyder, a geneticist at Stanford University in California and a member of ENCODE. "Most people would say between 10% and 20% of the [non-coding] genome is likely to have a function where, if you disrupt it, you will affect something."

But regulatory elements have a bewildering array of functions and forms, which makes tackling them a formidable challenge. Even the best-known types, such as spots in the genome known as promoters, which lie next to a gene where transcription begins, and enhancers — regions that when bound by specific transcription factors alter the likelihood of a gene being read — are hard to study. In addition to the sheer number of these sites, estimated at 15 million, enhancers may be positioned thousands of base pairs away from the gene that they control. This makes it tough to predict where their target genes are located and what they do.

Thus far, ENCODE and Roadmap have offered up important clues, but the real proof that these predicted regulatory elements actually do something comes from a functional test. For genes, this mostly entails deleting them one

at a time and observing the consequences in a cell assay or animal model. This is less easy to do for the non-coding genome because many of the elements are redundant, and so deleting just one might not alter gene expression or produce an obvious change. "It's a huge challenge that we have at the moment to really distinguish between functional and non-functional elements detected by ENCODE," says geneticist Ran Elkon of Tel Aviv University in Israel.

CRISPR–Cas9 is particularly accelerating scientists' exploration of enhancers. The technology enables scientists to alter large numbers of regulatory elements in a high-throughput way, using libraries of RNA guide fragments that target and disrupt different regions in the genome, to observe the outcome. Not only is the method relatively fast, but researchers can also run the assays directly in human cells.

Experiments of this type have already turned up some unexpected findings. While a postdoc working with cancer biologist Reuven Agami at the Netherlands Cancer Institute in Amsterdam, Elkon was involved in performing the first screen of regulatory elements using the advanced editing system<sup>4</sup>. The CRISPR–Cas9 approach enabled them to test individually those enhancers predicted by ENCODE to bind a transcription factor called p53. Interest in p53 is high because the protein is a known tumour suppressor that is mutated in more than 50% of human tumours. The researchers were able to pinpoint two enhancers from more than a thousand genomic sites that affect p53's tumour-suppressing function, located near the p53-encoding gene. A predicted third enhancer has yet to be located because it is far from any gene, let alone one related to p53.

In a separate screen, the group targeted binding sites for oestrogen receptor- $\alpha$  — which is implicated in breast cancer — and identified three enhancer sequences that influence tumour growth; these elements could thus have a role in the development of resistance to breast-cancer therapy.

At the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, bio-engineer Feng Zhang and his group also used CRISPR–Cas9 to identify genes essential to the survival of cancer cells. Using a melanoma model, they first screened around 18,000 genes in human cells to pinpoint ones that might underlie resistance to the melanoma drug vemurafenib. Then, in a follow-up study published last month<sup>5</sup>, they described a new screen that identified regulatory regions on either side of several resistance genes. Their findings fit well with ENCODE data that predict regulatory regions at these locations — and they also reveal new functional elements, says molecular biologist Neville

Sanjana, who conducted the research as a former postdoc in Zhang's group and now works at the New York Genome Center at New York University.

Other CRISPR–Cas9 screening data have challenged ENCODE predictions. Richard Sherwood of Harvard Medical School in Boston, Massachusetts, and his collaborators created an approach called a multiplexed editing regulatory assay<sup>6</sup> to screen for non-coding regions that might influence gene expression in well-known mouse embryonic stem-cell lines. Using this technique, they obtained quantitative information about the extent to which these regulatory regions might contribute to gene expression. Some of their results are discordant with regions flagged by ENCODE as potential enhancers because, when mutated, these areas did not affect gene expression.

Moreover, the researchers also discovered mysterious sections that they dubbed 'unmarked regulatory elements', or UREs, that do not fit into any category of functional elements. The team is currently exploring how widespread these UREs might be in the genome. This new type of assay, along with other gene-editing-based screens, will play an increasingly important part in the validation of ENCODE candidates, says Sherwood.

### TECHNIQUE TWEAKS

Investigators working on the ENCODE and Roadmap projects have relied mostly on a biochemical technique called DNase-seq, which sequences and maps all exposed regions of the genome. In these sections, the DNA is relaxed instead of tightly coiled around histones, and thus is more likely to facilitate transcription-factor binding that drives gene activation. By mapping these areas, investigators can pinpoint candidate enhancers, promoters, silencers, insulators and other regulatory elements in the non-coding genome (see 'Spot the regulators').

Another method, ATAC-seq, detects and sequences sites in the chromatin that are accessible to the transposase enzymes used for the assay. Both DNase-seq and ATAC-seq produce a genome-wide view of regions of open chromatin. According to the researchers, because such epigenomic profiles can map the extent to which genes are activated in certain cell types, they could be useful for clinical decision-making, and ATAC-seq is fast enough for this purpose<sup>7</sup>. Many, however, consider a technique known as chromatin immunoprecipitation (ChIP)-seq to be the most reliable for this purpose because it is the only one that can identify all potential binding sites for a given transcription factor.

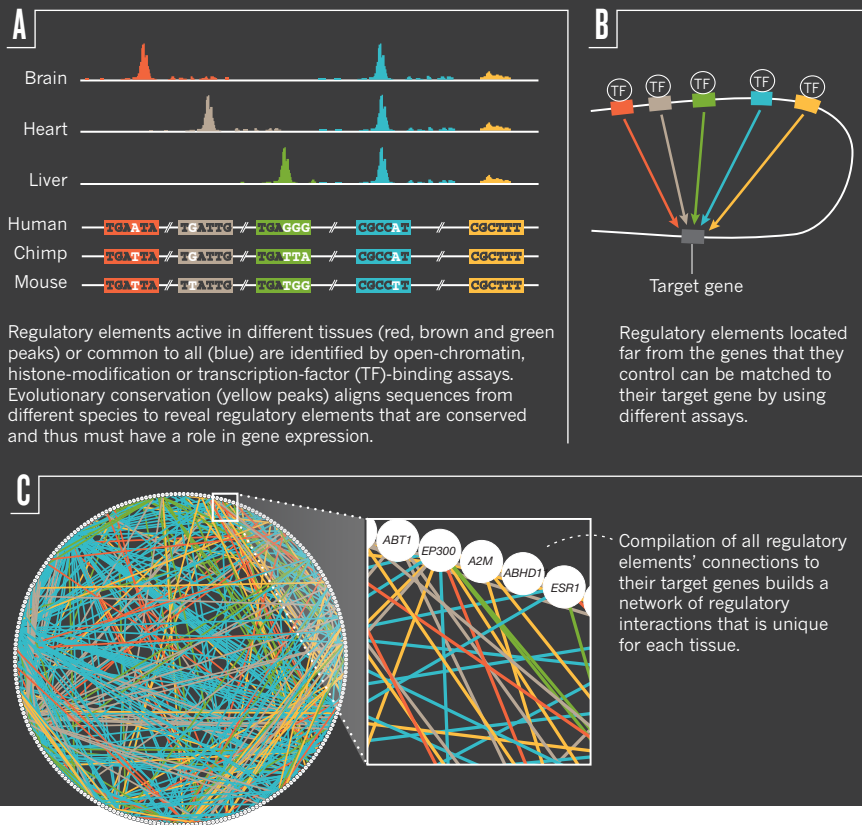
Even so, biochemical assays can only hint at function. CRISPR–Cas9 cell screens, by contrast, are more concrete because scientists can introduce a mutation or deletion at a particular site in the genome and observe how it influences gene expression. The disadvantage is that these

**"It's a huge challenge to distinguish between functional and non-functional elements."**



## SPOT THE REGULATORS

Scientists can identify functional regions in the DNA that are active in modulating gene expression by combining results from biochemical assays with evolutionary comparisons between species.



that is superimposed on a cell, and the result of that program is the development of genetic and genomic instability," says Stamatoyannopoulos. "As we've analysed lots of cancer genomes, all of these patterns now are starting to come out that were previously not imagined to exist."

It's possible that there are still elements in the genome that existing assays have missed. After all, regulatory signals still crop up unexpectedly, such as the UREs in Sherwood's screen. And a team of scientists led by Harvard Medical School immunologist Daniel Tenen discovered<sup>10</sup> a potential new class of regulators that seem to control whether a gene is turned on or off by blocking the enzyme DNA methyltransferase 1, which adds methyl groups to silence genes. These elements are dubbed 'extracoding RNAs', and because they can influence silencing in a gene-specific way, have therapeutic potential. Earlier this year, neuroscientist Jeremy Day of the University of Alabama at Birmingham and his colleagues showed in rat neurons that an extracoding RNA influences the transcription of a gene important for memory formation<sup>11</sup>.

The ENCODE team will continue to map the non-coding space in the genome and expects to cover most of the regulatory DNA by 2020, Stamatoyannopoulos says. A spatial understanding of how DNA is packaged into a cell, and of the 3D folding that positions genes in close contact with their regulatory elements, will be key to predicting an element's target genes. The NIH Common Fund has begun the '4D Nucleome' project, for instance, which aims to predict the target genes for every regulatory element. That knowledge will help to fill in the picture of how a given regulatory element influences health and disease.

Next-generation sequencing has been — and still is — the technological engine of ENCODE. But looking ahead, researchers might be able to roll out high-resolution live-cell imaging on a large scale to watch the state of the genome change in real time using specific markers. This technology could be disruptive. "If we had a better microscope, we wouldn't be sequencing anymore," says Stamatoyannopoulos. ■

**Kelly Rae Chi** is a freelance science writer based in Cary, North Carolina.

1. Ezkurdia, I. et al. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).
2. ENCODE Project Consortium *Nature* **489**, 57–74 (2012).
3. Bernstein, B. E. et al. *Nature Biotechnol.* **28**, 1045–1048 (2010).
4. Korkmaz, G. et al. *Nature Biotechnol.* **34**, 192–198 (2016).
5. Sanjana, N. E. et al. *Science* **353**, 1545–1549 (2016).
6. Rajagopal, N. et al. *Nature Biotechnol.* **34**, 167–174 (2016).
7. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. *Nature Meth.* **10**, 1213–1218 (2013).
8. Lee, D. et al. *Nature Genet.* **47**, 955–961 (2015).
9. Polak, P. et al. *Nature* **518**, 360–364 (2015).
10. Di Ruscio, A. et al. *Nature* **503**, 371–376 (2013).
11. Savell, K. E. et al. *Nature Commun.* **7**, 12091 (2016).

tests cover smaller portions of the genome. If the full genome of 3 billion base pairs were represented, for example, by three copies of Leo Tolstoy's classic novel *War and Peace* (1869), such screens would barely cover a single page, Sanjana says — although he is optimistic that future gene-editing approaches will scale this up.

"In the short term, I think CRISPR will serve mainly as a tool to validate functions predicted by those biochemical signatures," says Ren. Once enough of these kinds of screens have been done, their data could be fed into a machine-learning tool to improve its predictive power, Sherwood says.

New computational tools are already providing scientists with smart ways to interpret biochemical mapping data. Algorithms can predict transcription-factor binding sites, which researchers can then probe for function. But even with algorithms, predicting which enhancers are active in a given context is harder in human genomes than in yeast or worm genomes, says computational biologist Michael Beer of Johns Hopkins University in Baltimore, Maryland.

Beer and his collaborators have developed a computational model<sup>8</sup> to predict which tissue-specific networks of gene-regulatory elements are operating in a given cell type and to what extent they are perturbed in complex diseases. They trained their open-source algorithm,

called deltaSVM, on human lymphoblastoid cell lines using gene data from ENCODE in 2012, followed by mouse ENCODE data in 2014.

Scientists have initially focused on cancer to probe the links between functional elements and disease because cancer is a simpler condition to study at the cell level than, say, a neuropsychiatric disorder — cancer cell lines reveal simple-to-measure outcomes, such as cell multiplication, death or senescence. But the data that have streamed in from the Epigenome Roadmap consortium are shifting scientists' thinking about how cancers arise. A study published last year by geneticist John Stamatoyannopoulos of the University of Washington in Seattle and his collaborators showed<sup>9</sup> that mutations in a given cancer cell type cluster in inaccessible chromatin regions rather than in the exposed ones — possibly because the open regions can be accessed by DNA-repair enzymes.

The scientists also found that mutation density in a tumour is defined by the epigenomic profile specific to each type of cell. Consequently, the DNA sequence can be informative about tumour origin, which ushers in the possibility of using epigenomic data to trace cancer provenance in patients for whom it remains unknown. It could also open up new approaches to cancer treatment. "Cancer is essentially a regulatory or epigenetic program

# CAREERS

**INDUSTRY** Moving from an academic lab to a company role **p.281**

**STUDY ABROAD** An Indian scientist in Germany [go.nature.com/2csidtn](http://go.nature.com/2csidtn)

**NATUREJOBS** For the latest career listings and advice [www.naturejobs.com](http://www.naturejobs.com)

ARTQU/GETTY



**VOLUNTEERING**

## Pro bono science

*Offering your skills as a volunteer can be rewarding but needs careful thought.*

**BY AMBER DANCE**

**W**hen physician Alison Smith retired in 2012, she wasn't ready to hang up her stethoscope. So the primary-care doctor in Sussex, UK, decided to volunteer for the charity Freedom from Torture in London. She examines asylum seekers who need medical evidence of their abuse to strengthen their applications to stay in the country.

In part, the work helps to assuage what she describes as "survivor guilt" — the feeling that whereas she's had a great life, others haven't been so lucky. It also makes her feel that she's counterbalancing the acts of physicians who participate in torture. Plus, Smith admits, she has a selfish reason. "It's an interesting challenge," she says. "I don't want to let my professional skills go."

Smith is one of many scientists, engineers

and health workers worldwide who apply their professional abilities and skills — and their often-scant spare time — to volunteer missions. Not only does it give them satisfaction to deploy their skills for good, it can also offer career benefits such as opportunities for networking or, occasionally, for publishing. Researchers who feel a pull towards philanthropy can contribute to projects that involve rights to health or to safe living environments, for example. And they need not wait until retirement or travel abroad to get involved: there are ways to make a positive difference in the lives of many with a local commitment of a few hours. But it is important for would-be volunteers to select their projects carefully — and to grasp what is required of them in scientific expertise, time commitment and costs they might have to shoulder.

How to find an organization to volunteer

with? Some researchers land projects through their universities, and many turn to their discipline's professional associations. The On-call Scientists programme, run by the American Association for the Advancement of Science (AAAS) in Washington DC, matches scientists with human-rights organizations. It has enrolled nearly 1,000 scientists from 58 countries; tasks might include helping organizations to analyse and understand technical reports, designing surveys or developing useful technologies. "I'm seeing more and more of these kinds of 'without borders' types of initiatives started by scientific organizations," says Theresa Harris, manager of the programme.

One is the Thriving Earth Exchange (TEX), launched in 2013 by the American Geophysical Union in Washington DC. It matches Earth scientists with communities that need ►



► advice, such as cities that are planning for the effects of climate change. Natasha Udu-gama, director of community partnerships for TEX, notes that scientists often begin their careers eager to help humankind. Then they get sidetracked by the pressure to publish, write grants, achieve tenure and reach other career milestones. Humanitarian projects, even small ones, can help to fulfil that lost-along-the-way desire to make the world better.

It's the feeling that inspired Linus Bengtsson, an epidemiologist at the Karolinska Institute in Stockholm, to found a non-profit humanitarian organization. "Everybody who deals with global health, they want to accomplish something in the end, and many feel that research is not enough," he says. In 2008 he co-founded Flowminder, which uses data about people's locations — from surveys, satellites and mobile phones — to map where people are in low- and middle-income countries. These data can help charities and governments to direct aid or to track transmission of disease. About half of the people who work with the organization are volunteer scientists who may incorporate the studies into their own research; the rest are staff.

### IN DEMAND

Some types of expertise are much in demand, including statisticians to crunch numbers, environmental scientists to evaluate reports on hazards, and physicians to provide medical assessments or care. But scientists and engineers can also contribute to projects outside their specific fields. "There are so many transferable skills," Harris says. "Anybody who knows how to clean up a database is valuable."

Perhaps one of the most attractive elements of volunteer work for researchers is that they can make a huge difference in others' lives more rapidly than with most research projects. A member of On-call Scientists, for example, helped the small town of Temacapulín in Mexico as it faced the danger of flooding as a result of the building of a dam. Harris put the town's lawyers in touch with a water-resources engineer who analysed the proposed dam, attended public hearings and helped to develop a plan to scale down the dam and safeguard the town. Statistics Without Borders, an outreach committee of the American Statistical Association in Alexandria, Virginia, provided volunteers to help Nepal to reorganize its census data in the wake of the April 2015 earthquake. The new spreadsheet allowed fieldworkers to know who was living in the communities they visited. And DataKind, headquartered in New York City, a global network of volunteers skilled in working with data, assisted US-based Crisis Text Line, a helpline for teens. They created an algorithm to identify the most urgent texts and prioritize responses.

For researchers who have time to travel, or can combine philanthropy with work trips, off-site volunteering is an option. Mary Gray, a statistician and lawyer at the American University in Washington DC, spent two weeks in



Volunteers helped to track survivors of a 2015 earthquake in Nepal.

Fiji in 2011 when local hospital officials wanted to assess whether their medical care was effective. Gray gave doctors and lawyers a crash course in statistics, which helped them to make those judgements. "In a week or in two you can accomplish quite a lot," she says. "You leave a resource behind."

Those who can't travel can often find opportunities close to home. Climate scientist Andrew Rice at Portland State University in Oregon helped the nearby city of Eugene with its carbon-emissions-cutting plan in a TEX project. He examined the approach and the maths, provided by a city official and a consulting firm, on how much of the city's carbon emissions would have to be cut. He attended a city-council

meeting, where he spoke about how the team came up with its figures, and how the numbers related to global climate-change projections. The city approved an emissions-cutting target based on the project in July. Rice liked knowing that his scientific skills had immediate impact. And Udu-gama of TEX says that cities appreciate input from an unpaid third-party expert.

Volunteering can also expand one's skill set. Volunteers get a taste of project management and learn to work with clients from different cultural backgrounds. Teamwork is especially crucial in pro bono projects. "You must be willing to listen," says Udu-gama. "We stress that above all else."

### CAREER BOOST

Volunteer work can also help a career. Junior scientists and students may gain experience that they can add to their CVs, and researchers may be able to publish the work they did or integrate it into their current projects. Ashley Collier, a graduate student in environmental engineering at the University of Colorado Boulder, plans to include her volunteer work in her thesis on low-cost air-quality monitoring. Through TEX, she got involved with a community group in Denver, Colorado, that was concerned about contamination from perchloroethylene, a chemical used in dry cleaning. Collier and collaborators tested homes for perchloroethylene as well as radon gas. They found no worrying levels of perchloroethylene, but discovered excessive radon in 12 out of 15 homes. Now, Collier and the community group Taking Neighborhood Health to Heart are planning ways to educate residents about radon and how they can fix their homes to protect themselves.

## VOLUNTEER PREPARATION

### Questions to ask

When seeking or preparing for volunteer work, it is important to communicate with the charity about both its expectations and yours, and set the plan down clearly, possibly even in a contract. "Treat your pro bono engagement as you would a paid engagement," advises Miriam Young, head of communications and culture for the data-driven non-profit DataKind in New York City.

Here are some questions to ask the non-profit, and yourself, before signing on the dotted line.

- Is this a credible human-rights organization? What are its goals?
- How much time will I be spending on this project? How urgent is it, and how long will it take?
- What are my deliverables?
- What's the budget? What expenses will I have to pay?

- Are there confidential data to protect, and who will have access?
- Do I need to do anything to make sure findings are admissible in court?
- Will there be any conflict of interest with my current or past employment?
- Do I need any insurance, such as liability or travel coverage?
- How will the non-profit use these results?
- Who owns the samples and data? Can I publish the results?
- Will I have to testify in court? Or talk to the media?
- How will the organization take care of my well-being?

For more guidelines, see *Human Rights Projects: Guidelines for Scientists and Human Rights Organizations* (American Association for the Advancement of Science, 2012), available at [go.nature.com/2d7dtgx](http://go.nature.com/2d7dtgx). **A.D.**

Scientists working with Flowminder can adjust their research to better fit people's needs, Bengtsson says. Flowminder researchers collaborated with a Nepali mobile-phone operator to map how people moved around after the 2015 earthquake, producing data that were useful to relief organizations. In turn, the work led to a publication for the researchers (R. Wilson *et al.* *PLoS Curr. Disasters* <http://doi.org/brbp>; 2016).

Stéphanie Horion, a postdoc at the University of Copenhagen who studies global environmental change and Earth observation, found that a TEX project pointed her to a resource that she is now applying in her research. TEX asked her to help the Horn of Africa Regional Environmental Centre and Network, a group in Addis Ababa, Ethiopia, that promotes sustainable development and environmental issues. It wanted to understand how a region surrounding Lake Ziway was using water. It already had someone to build a model of the water usage and requirements, but needed additional satellite maps. Horion contacted her network of scientists to find the right maps. In the process she accessed new vegetation maps, which she is now using in her own analyses of ecosystem change.

Networking while volunteering can also be a significant career perk. Samarth Bhargava, a software engineer at Inferlytics in Bangalore, India, schmoozes with other data-heads at DataKind events. "You get to interact with these amazing people and pick their brains," he says. And in his DataKind project to help a non-governmental organization process web-based complaints and predict future problems (such as a spate of messages about overflowing drains that might herald floods when monsoons hit) he's learnt about urban planning and government operations.

### CAREFUL PREPARATION

There can be pitfalls, especially if the non-profits or the scientists who work with them don't communicate enough at the start (see 'Questions to ask'). Organizations may not understand the limitations of what scientists can provide if they are not spelled out. Harris recalls one non-profit that asked for an On-call Scientists volunteer to design a survey, which the scientist completed. But when the non-profit asked how the survey would be distributed, the scientist was not able to do this, and both sides were left disappointed.

Harris also warns that non-profits often have to reorganize their priorities in a hurry, pushing projects onto the back burner as a result of political changes in the host nation or emergencies such as a tsunami. Scientists should understand this and should also keep in mind that their proposed solution may not be used.

Volunteers may have to foot some expenses themselves. When Gray went to Fiji, she paid her own airfare. Smith estimates that she spends about £2,000 (US\$2,600) per year on her medical-doctor registration, liability insurance and travel to and from London to work with Freedom from Torture. And volunteers need to be realistic about how much time they can offer.

Some say it would be easier to volunteer for longer periods once one is retired or established in a career, but others say there's no reason not to start early. The summer months or between semesters, when academic responsibilities diminish, are good times. "People should at least be thinking about these things as a graduate student — what are the societal implications of your research?" says Rice.

The experience required varies by organization. For example, On-call Scientists seeks volunteers with at least a master's degree plus two years of experience in their field. Statistics Without Borders lets anyone sign up, no matter their level of education or experience. Foreign-language skills help to boost a volunteer's appeal, Harris notes. And some profes-

**"You get to interact with these amazing people and pick their brains."**

sors may be able to get their students involved in projects, thus enabling younger scientists to get experience in volunteering. Being accepted isn't the only hurdle to successful volunteering. Researchers should make sure that the organization or project will have the desired impact on people in need, says Bengtsson. At AAAS, Harris vets non-profits by checking their past accomplishments and speaking with referees, such as peers at other charitable organizations. Scientists can check a prospective partner by reviewing the organization's reports, and by checking websites such as Charity Navigator or GuideStar, which provide information on and evaluations of charities and non-profits.

After that, the choice is straightforward. "Do something that you think is interesting," Bengtsson advises. "Follow your heart." ■

*Amber Dance is a freelance writer in Los Angeles, California.*

### CLARIFICATION

The general advisory nature of the Careers article 'How to survive as a whistle-blower' (*Nature* **532**, 405; 2016) meant that it purposely did not discuss the case used as an example in detail. For clarity, it should be noted that as a result of the university's initial internal inquiry into the case, the co-authors were recommended to submit a list of errors to the journal in question. The editor-in-chief of the journal ultimately elected to retract the paper. In addition, a subsequent external inquiry concluded that although there had been misconduct by one researcher, there was no misconduct by the travelling supervisor referred to in the article, and there was no intentional wrongdoing. It also found that some of the allegations were not substantiated.

## TRADE TALK

### Lab whiz



*As a postdoc and graduate student, Stefanie Marquez-Vilender acquired a suite of lab skills. Now she works at AxoGen, a medical-technology firm in Alachua, Florida.*

#### How did you approach your job interview?

I looked at the job description and what skills they were looking for, and came up with a response to how I would fit. Even if I hadn't written regulatory documents, I could say I had written grants. It's better to say that than to say that you don't have any clue.

#### Why did your company choose you?

They wanted someone who could provide support for all sorts of experiments, and to help with writing protocols and with university collaborations. Even though I didn't have all the techniques, I had shown I could learn many new methods in my postdoc. And I did have management experience coordinating 10–15 people, mainly undergraduates. I work closely with students and interns here at AxoGen, helping them to plan what experiments they need to run and what reagents they need to buy.

#### Do you wish you'd done anything differently during your training?

For graduate school I found a lab that was interesting to me, rather than a lab that published a lot or had a lot of funding; those were questions I probably should have asked about when I was choosing. An important question to ask about potential mentors is how often they participate in conferences. As a postdoc, I didn't attend a single conference outside the University of Florida. This hurt my chances of getting a job quickly, because I didn't make any industry contacts.

#### How can trainees best prepare for their job search?

I don't regret doing a postdoc, but I wish that I had had different opportunities presented to me. We have quite a few interns here; they are getting experience that will make them more competitive if they decide to go into industry. I didn't have that available. It was that lack of industry experience that was holding me back — it was hard for me to get my foot in the door.

### INTERVIEW BY MONYA BAKER

This interview has been edited for length and clarity. See [go.nature.com/2dbh9u8](http://go.nature.com/2dbh9u8) for more.



# GALACTIC EMPIRE MANAGEMENT CONSULTING: LESSONS LEARNT

*Prepare for excellence!*

BY AIDAN DOYLE

**H**i everyone! I'm a senior consultant with Galactic Empire Management Consulting. I've been brought on board to analyse the previous project's failure and to help bring the new project back on schedule. We have a tight deadline, but I'm sure if we all work together we can achieve those key milestones! I look forward to working with all of you!

++++

I'd appreciate it if everyone took a few minutes to answer the attached survey questions. We're undertaking an Organizational Culture Inventory (OCI) of the Empire's workforce. It's widely recognized that workplace culture underpins everything we do, including how we approach our work and interact with each other. Improving workplace culture can help us innovate and succeed as an organization.

++++

After conducting a preliminary review, it looks as though the new project has failed to learn from previous mistakes. In our experience, when it comes to integrating new planets into an empire, building a planet-destroying space station is simply not cost efficient. Other clients have achieved greater results through building hospitals, schools and cultural centres. Our initial report also highlights key failures in process management and threat assessment. I trust you'll all make time to thoroughly review the document. Also, please note, as of Wednesday, the use of emoji will no longer be permitted in work-related messages.

++++

I've uncovered disturbing reports about staff being discriminated against due to their lack of faith. I'd like to remind everyone that religious discrimination in the workplace will absolutely not be tolerated. I'll be scheduling a series of mandatory sensitivity training workshops over the next couple of weeks.

++++

A happy workplace is a productive workplace! When one of your co-workers helps you, remember to let them know you appreciate their actions. You now have the opportunity to thank your colleagues by using the attached *Thank You Template!*

Not only will recipients love getting your thanks, but you also have the chance to win one of five cantina voucher rewards in a monthly draw!

++++

After reviewing a number of military supplier contracts and conducting an analysis of key battle performance reports, we've reached the conclusion that supplier claims about the accuracy of their weapons reducing the need for troop training have been exaggerated. As a first step to ensuring our armed forces maintain their famed combat excellence, we recommend a complete systems upgrade of the imperial personnel training database.

++++

It is with great disappointment that I inform you that members of the engineering division have been disciplined for posting criticisms of the project's revised work schedule on social media. The new schedule was developed with the input of leading business analysts and experienced project managers and is entirely achievable. I shouldn't have to remind staff that the new schedule has the Emperor's approval. I'll also take this opportunity to reiterate that staff need to seek management approval before posting on social media.

++++

I'd like to remind middle management that summary termination of staff must receive approval from the Executive Leadership Team. We don't want to risk violating workplace agreements that could render the Empire liable to compensation claims and industrial action.

++++

Please find attached an updated version of the imperial guide to using the tractor beam. This guide has been developed in consultation with the beam's manufacturer, along with workplace safety experts. Staff are reminded that using the tractor beam for personal reasons is strictly prohibited.

++++

An analysis of the imperial performance-based pay system has discovered the practice of rewarding staff with bonuses for subduing rebel systems has led to staff abusing this system. Some staff have been found guilty of encouraging newly integrated planets to embark on a rebellion, then reconquering them to claim the bonus multiple times.

Pending further review, this pay bonus has been discontinued.

++++

Our new values are here! I'm pleased to share the Empire's new values: Service Excellence, Ownership, Teamwork and Wellbeing. These have been developed by the Executive Leadership Team based on feedback provided by staff. These values will help clarify our mission to bring peace and prosperity to all of the Galaxy. I encourage you all to reflect on our new values and consider how you can bring them to life through your behaviours. We will be asking for your thoughts and input on behaviours over the coming weeks.

++++

Engineering staff are reminded to regularly change their password. It is vitally important that classified documents do not fall into the wrong hands.

++++

The Empire has raised the Imperial Terrorism Public Alert Level from medium to high. In response to this heightened security alert level, the Executive Leadership Team has mandated the use of colour-coded security badges for all staff, contractors and visitors entering imperial work bases. The badges are coded as follows: blue for members of the imperial armed forces, red for contractors and bounty hunters, and yellow for visitors. Thank you for your cooperation.

++++

Your feedback is requested on the three finalists in the Design a New Imperial Logo contest. The winner will have the honour of seeing their design emblazoned on imperial spacecraft and military uniforms.

++++

Please find attached the final report on the failures of the previous project. In my experience, these reports are regrettably often ignored, but I hope you will take the time to thoroughly review it and implement its recommendations. The fate of the Galaxy is in your hands! Thank you for all your support and have a great day! ■

*Aidan Doyle is an Australian writer and computer programmer. He has visited more than 90 countries and his experiences include teaching English in Japan, interviewing ninjas in Bolivia and going ten-pin bowling in North Korea.*

ILLUSTRATION BY JACEY